EVALUATION OF FACTORS AFFECTING PERFORMANCE OF DIAGNOSTIC TESTS FOR INFECTIOUS SALMON ANAEMIA VIRUS

A Thesis

Submitted to the Graduate Faculty

in Partial Fulfillment of the Requirements

for the Degree of

Doctor in Philosophy

in the Department of Health Management

Faculty of Veterinary Medicine

University of Prince Edward Island

Charles G. B. Caraguel

Charlottetown, P. E. I.

February, 2010

PERMISSION TO USE POSTGRADUATE THESIS

Title of thesis: Evaluation of factors affecting performance of diagnostic tests for infectious salmon anaemia virus

Name of Author: Charles G. B. Caraguel

Department: Health Management

Degree: Doctor in Philosophy Year: 2010

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from the University of Prince Edward Island, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the professor or professors who supervised my thesis work, or, in their absence, by the Chair of the Department or the Dean of the Faculty in which my thesis work was done. It is understood any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Prince Edward Island in any scholarly use which may be made of any material in my thesis.

Signature:

Address: AVC-Centre for Aquatic Health Sciences & Department of Health Management
Atlantic Veterinary College
University of Prince Edward Island
Charlottetown, Prince Edward Island, Canada
C1A 4P3

Date: February 2010

University of Prince Edward Island

Faculty of Veterinary Medicine

Charlottetown

CERTIFICATION OF THESIS WORK

We, the undersigned, certify that **Charles G. B. Caraguel**, candidate for the degree of **Doctor in Philosophy** has presented his thesis with the following title: **EVALUATION OF FACTORS AFFECTING PERFORMANCE OF DIAGNOSTIC TESTS FOR INFECTIOUS SALMON ANAEMIA VIRUS** that the thesis is acceptable in form and content, and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate through an oral examination held on February 12th, 2010.

Examiners: Dr. Javier Sanchez (Chair) _____

Dr. Ian Gardner (External) _____

Dr. Peter Wright _____

Dr. Crawford Revie _____

Dr. Larry Hammell _____

Date: _____

ABSTRACT

To secure the safety of international trade of animal and their derived products, it is required that animals should be proven infection-free using a validated diagnostic test certified by the World Organization for Animal Health (OIE). Test validation involves a multi-step evaluation process to assess test accuracy and its fitness for a specific purpose. Contrary to common understanding, diagnostic accuracy is not constant for each test and may vary within and between populations according to the distribution of biological factors that influence the pathophysiology of the disease. Diagnostic test accuracy combines the *precision* (repeatability & reproducibility) and the *trueness* of the test (diagnostic sensitivity, DSe, & specificity, DSp).

The objective of this research program was to extend diagnostic test evaluation methods by estimating accuracy specifically for influential covariate factors. This research was applied to reverse-transcriptase polymerase chain reaction (RT-PCR) for infectious salmon anaemia virus (ISAV). Early detection of ISAV is the base of an efficient control of this devastating disease for salmon aquaculture industries.

RT-PCR accuracy was shown to differ among stages of infection, revealing substantial variation of test precision across prevalences of infection stages in the tested populations. Submission factors such as homogenization and testing laboratory also significantly impacted the precison and were accounted for to predict test result agreement across prevalences. Latent Class Modeling (LCM) was used to evaluate test trueness in absence of true status information. LCM assumes that DSe & DSp are constant across populations, which was revealed as an invalid assumption. Extending to 3-class LCM revealed different DSe between low- and high-infected salmon. Finally, the selection of the proper cutpoint for the real-time version of RT-PCR was dependent on the distribution of infection stages in the target population. Both analytical and epidemiological approaches to select the cutpoint were reviewed to improve the fit for intended purpose of the test.

Potential applications for specific estimates of test accuracy and new perspectives on diagnostic test evaluation and use were discussed.

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to my co-supervisors Drs. Larry Hammell and Ian Dohoo. Primarily to Larry for believing in me and providing me with all the financial and logistic resources necessary to satisfy my endless needs. But most of all, because Larry was, and still is, an inspiring mentor with a unique instinct and vision for research. I believe this program was the beginning of a long collaboration between us in a domain where both of us share a passion. Thanks also to Ian for his extraordinary teaching skills, and for providing me with the best possible training by consistently challenging my thoughts and pushing me further in my reasoning. I am no exception when I say, Ian represents a model that several generations of veterinary epidemiologists want to reach, and maybe exceed, but the bar is high and ever rising. In addition, I would like to extend my gratitude to others members of my supervisory and examination committees: Drs. Peter Wright, Dave Groman, John VanLeeuwen, Crawford Revie, Ian Gardner and Javier Sanchez.

Special thanks to Dr. Henrik Stryhn who was instrumental from the philosophy to the achievement of this work. The quality of this thesis was significantly levered by his collaboration. He is one of the rare people who does not fear my questions, but from whom I fear the answer. Thank you also to Nellie Gagné for her immoderate support and synergetic collaboration during this project. I am thankful to Dr. Frank Berthe that convinced me to pursue a PhD program, to Dr. Spencer Greenwood who encourage me to expand my horizons, and to Dr. Pascale Nérette who inspired me to explore diagnostic test investigation.

I wish to thank the technical staff and casual employees of the AVC-Centre for Aquatic Health Sciences for the impressive effort and efficacy provided during sample collection events. Also, all the science and management staff for their enthusiasm and team spirit, in detail: Drs. Carol McClure, Jillian Westcott, Shona White, Andrea McKenna and Holly Burnley.

Thanks to all the people that have participated in my PEI life for the last 6 years:
-all my fellow graduate students, faculty and staff from the various departments and services,
- Peter Sykes and Kathryn Evans, and their respective families, who welcomed me as one of them, and whom I admire for their altruism and much more,
- the various sports communities I have been involved with (too many to be cited)
- Beatrice Després and the french speakers connection for keeping me rooted.

I am grateful to my family and friends that tolerated the distance that separate us and the lack of contacts and still support me in my professional choices even though they do not understand the word “epidemiology”.

Finally, I deeply thank Kyle for supporting me and helping me to stand when the days were dark and cold.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ASe	Analytical sensitivity
ASp	Analytical specificity
df	Degrees of freedom
DSe	Diagnostic sensitivity
DSp	Diagnostic specificity
ELISA	Enzyme-Linked Immunosorbent Assay
EQC	External quality control
FAO	Food & Agriculture Organization
HPR0	Highly polymorphic Region-0 (genotype of the low-virulent strain of ISAV)
IFAT	Indirect Immunofluorescent Antibody Test
IQC	Internal quality control
ISA	Infectious salmon anaemia
ISAV	Infectious salmon anaemia virus
LCM	latent class model
LFI	Lateral Flow Immunoassay
LOD	Limit of detection
LR	Likelihood ratio
LR+	Likelihood ratio of a positive test result
LR-	Likelihood ratio of a negative test result
NPV	Negative predictive value
NPV+	Negative predictive value of positive test result
NPV-	Negative predictive value of negative test result
OIE	World Organisation of Animal Health (Office International des Epizooties)
Pa	Proportion agreement
Pa+	Positive agreement
Pa-	Negative agreement
PCR	Polymerase Chain Reaction
PPV	Positive predictive value
PPV+	Positive predictive value of positive test result

PPV-	Positive predictive value of negative test result
Pr	Prevalence
QA	Quality assurance
QC	Quality control
qPCR	real-time Polymerase Chain Reaction
qRT-PCR	real-time Reverse-transcriptase Polymerase Chain Reaction
RT-PCR	Reverse-transcriptase Polymerase Chain Reaction
r	Repeatability
R	Reproducibility
US	United States of America
VI	Virus isolation

Chapter I: GENERAL INTRODUCTION

1.1 Introduction

In 2005 world-wide, the contribution of fish to total animal protein intake was 15.3% which represents 16.4 kg per capita. In low-income, food-deficit countries, fish protein contribution was significantly higher with 18.5% (13.8 kg per capita). From the 143 million tonnes of fish produced in 2005, approximately a third was supplied by aquaculture. While the global capture fisheries production has been stable over the last decade at around 90 million tonnes per year, aquaculture continues to increase the fish supply and represents the fastest growing food animal production, outpacing population growth (FAO, 2009). In terms of consumption, a recent study reported that aquaculture provides half of the fish protein consumed (Naylor et al., 2009).

Development and implementation of stock health programs represents the cornerstone of sustainable aquaculture. Fish health management is further hampered by limited pathognomonic clinical manifestations of disease. The intensive international exchange of fish products represents increased risk of disease introduction and dissemination, resulting in direct economic, food safety and environmental impacts. Therefore, use of standard diagnostic tests remains an important method of evaluation of the health of aquatic animals, and their products, at the local, regional, national and international level.

Etymologically, the adjective “diagnostic” is derived from the ancient Greek root of diagnōsis (διάγνωσις), meaning dia- “split”, and gnosi “to learn-knowledge”. It refers

to the capacity to discern or distinguish a disease by its signs, symptoms, and from the results of testing procedures. In this thesis, the terms "test", "assay" or "test method" will be employed as synonyms and refer to the principles, systematic procedures, and processes used to detect or quantify an analyte of interest. International institutions and local authorities are mainly concerned with infectious diseases that can be introduced and transmitted to populations free of the pathogen. Thus, in the usual context, "diagnostic test" is assumed to refer to the technique that identifies an infectious disease of interest, and more specifically that detects infection and/or carrier stages that are frequently asymptomatic (i.e. presence or absence of the infectious agent). It is possible to classify assay outcomes as continuous, ordinal or dichotomous. This introductory chapter will focus on the evaluation of **dichotomous test results** (i.e. binary outcome: positive or negative) since, often, the two other test categories (i.e. continuous and ordinal) are dichotomized for decision making using a threshold or cut-point value. Also, the terms "positive" and "negative" are used here to characterize a test result, while "non-infected/non-diseased" (D-) and "infected/diseased" (D+) will be used to describe the status of an individual or a sample.

The interpretation of diagnostic tests results is particularly pertinent for disease surveillance and control in the case of infections where therapeutic options are lacking. A well-suited disease for investigating the use of diagnostic tests for surveillance and control in aquaculture is infectious salmon anaemia virus (ISAV) which represents a serious economic threat to the Atlantic salmon farming industry worldwide. This thesis addresses, develops and applies diagnostic test evaluation methods within an ISAV framework. The introduction chapter reviews the international and national requirements

to evaluate diagnostic test accuracy, the estimation methods and their limitations, and the application of these estimations.

1.2 International and national structure for diagnostic test validation

Aquaculture has in recent years has become globalized and is now primarily ruled by international trade and production policies. Historically, the Office International des Epizooties (OIE) set diagnostic requirements exclusively for international trade or movement of animals and their products (Wright et al., 2006). This organization, which has been recently renamed as the World Organization of Animal Health, progressively expanded its scope to include guidelines for the integration of diagnostic applications in the development of health programs in local settings (Wright et al., 2006). As a consequence of these internationally agreed upon protocols, standards for diagnostic tests used in aquatic species primarily follow international guidelines before being adapted into national or regional programs.

1.2.1 World Organization for Animal Health (Office International des Epizooties)

The Office International des Epizooties was originally created in 1924 in Paris from an international agreement to address animal disease (OIE website). In 2003, it was renamed the World Organization for Animal Health but still uses the same well-recognized acronym (OIE) known world-wide. The OIE membership is comprised of 174 countries and territories and provides leadership in improving animal health worldwide

(OIE website). Every year, the OIE publishes the "Health Code for Aquatic Animals" (OIE, 2009a) that establishes the international standards for trade of aquatic animals and their products. The code lists infectious diseases of concern for international trade and refers to methods for testing and self-certifying animals as specifically disease-free before product movement is allowed across national boundaries. Detailed protocols for **prescribed tests** are published in parallel in the "Manual of Diagnostic Tests and Vaccines for Aquatic Animals" (OIE, 2009b). The manual also outlines protocols for **alternative tests** that can be used within local settings or for import/export after bilateral agreements are established (Wright et al., 2006). To be included in the manual as a prescribed or alternative test, a new diagnostic method must be validated and approved by the OIE's Biological Standards Commission. Outlined in detail in the introductory chapters of the manual, the OIE also provides condensed guidelines on the validation of tests in the "OIE Standard for Management and Technical Requirement for Laboratories Conducting Tests for Infectious Animal Diseases" (OIE, 2008). This standard is an interpretation of the international quality standard for testing laboratories (ISO/IEC 17025: 2005) with a clear conceptual distinction that the validation of a test is appropriate only if it is relevant for its intended application (Wright et al., 2006). In other words, the test must be "fit for purpose" and validated as such (Wright et al., 2006).

The overall evaluation process begins with the actual development of the assay, followed by validation and finally continuous monitoring of the assay performance once the assay has been put into routine diagnostic service (Fig. 1.1). These three phases in the process are independent but equally important for the OIE quality standards. According to the OIE guidelines, test validation is an incremental process composed of 4 stages.

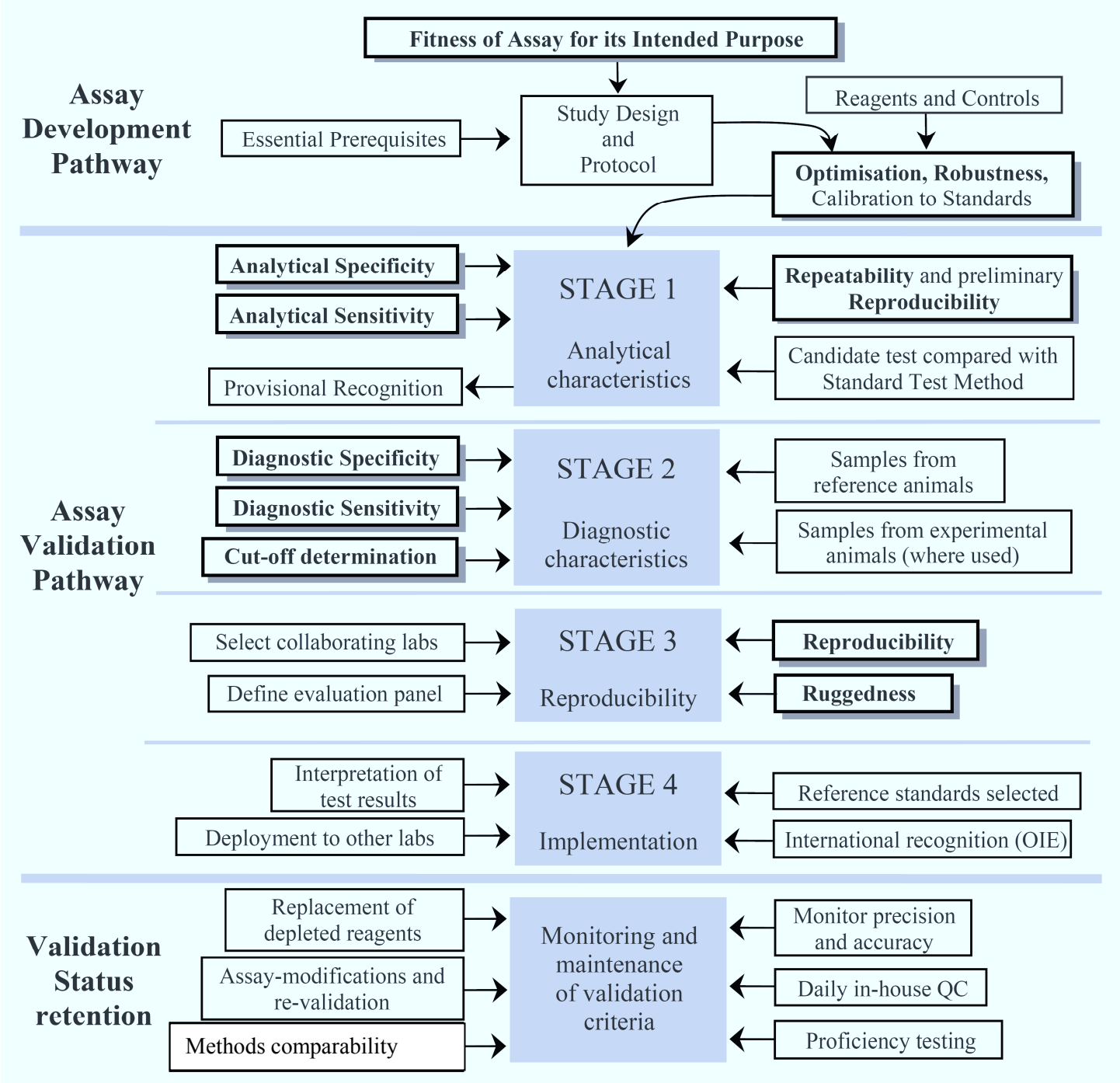


Figure 1.1. Assay development and validation pathway according to OIE guidelines (adapted from OIE, 2009b). Assay validation criteria of interest are in bold.

1.2.1.1 *Assay development pathway*

While provisional status may be recognized after the first stage (Fig. 1.1) of validation, data to support the next validation stages, collected over a period of actual use, is essential for full recognition. In the following sections, the three major pathways consisting of development, validation and maintenance as reported in the OIE manual (OIE, 2009b) and the assay validation criteria required for certification will be reviewed.

Prior to the development of a new assay, the method designers must clearly define the intended purpose of the test and the laboratory has to fulfill "essential prerequisites". OIE classifies **diagnostic test purposes and applications** into six general categories:

1. *Demonstration of freedom from infection (prevalence apparently nil) in a defined population including 3 situations: 'free' with vaccination, historical 'freedom', re-establishment of 'freedom' post-outbreak*
2. *Certification of freedom from infection or agent in individual animals or products for trade purposes*
3. *Eradication of disease or elimination of infection from defined populations*
4. *Confirmatory diagnosis of suspect or clinical cases*
5. *Estimation of infection or exposure prevalence to facilitate risk analysis (surveys, classification of herd health status, implementation of disease control measures)*
6. *Determination of immune status in individual animals or populations (post-vaccination)*

According to the purpose, the validation requirements for operating characteristics of the test may be different. For instance, if a test is intended for *screening* for an infectious disease (e.g. category #1 above), its diagnostic sensitivity will be prioritized, but when intended for disease *confirmation* (e.g. category #4 above), its diagnostic specificity will be prioritized (see section 1.2.3.7). In addition, the facilities where the test will be developed require that a system of quality assurance (QA) and quality control (QC) be established to ensure high value and confidence in the results.

The technical development of a test method includes the design of a detailed protocol describing the selection of reagents and reference materials for controls. Biological, physical and chemical parameters must be optimized to suit the intended test purpose. First, the influence of biological factors, such as the analyte concentration, is assessed by estimating the **linear operating range of the assay**. Defined as the interval of analyte concentration over which the test provides suitable accuracy and precision (OIE, 2009b), the linear operating range describes the "dose-response curve" and implies preliminary estimation of the lower and upper limits of detection of the assay. Then, the **robustness** is evaluated to investigate the critical physical parameters that may impact test performance. Also defined as the resistance of the test to expected variations of testing conditions (OIE, 2009b), robustness identifies critical factors during transportation, storage and repeated use that may impact the stability of reagents (Crowther et al., 2006). A test method may be "transport robust" and/or "laboratory robust" (storage and handling). The objective of this assessment is ultimately to define the optimal transport and laboratory conditions in which the test would routinely be used to ensure consistent operating characteristics.

The last step of the development pathway is the calibration of the test method. This aspect of standardization requires use of reference samples (or standards), well defined in terms of analyte and concentration. To facilitate calibration and comparison of results among testing laboratories, international and/or national reference standards should be used, where available. These highly characterized standards are usually available for international reference laboratories in limited quantity as *primary reference standards* or may be prepared by national reference laboratories as *secondary and/or working standards*. As a *working (or tertiary) standard*, they are used to monitor and control the quality of test results on a continual run-by-run basis. They may also be used to normalise test results of a continuous nature, resulting in units that are standardized across laboratories (e.g. optical density ratios for ELISA's).

At this stage, the test method should be robust and functional and available for further evaluations of its analytical and diagnostic performance through the validation pathway.

1.2.1.2 Assay validation pathway

By OIE definition, the validation of a test refers to the "process that determines the fitness of an assay that has been properly developed, optimized and standardized for an intended purpose". Illustrated in Fig. 1.1, the validation pathway is a succession of 4 evaluation stages: analytical characteristics, diagnostic characteristics, reproducibility and implementation. Although the assay can receive a provisional recognition after the analytical evaluation (Stage 1), the method will only be considered fully validated for the

original intended purpose if the first three stages of the validation pathway are completed. According to Van den Bruel et al. (2006), the procedure prescribed by OIE would only fit the studied population (*internal validation*). The authors further suggested that an additional confirmatory study be conducted on an independent but similar population to externally validate the results of the first evaluation (*external validation*). This is the intent of the last stage 4 described below (see section 1.2.1.2.4).

1.2.1.2.1 Stage 1- Analytical performance characteristics

The first analytical validation criterion of interest is the **repeatability**. According to the international standards, repeatability is defined as a measure of variation in test results that are obtained using the same method on identical test items in the same laboratory by the same operator with the same equipment over a short interval of time (within-laboratory consistency) (IS0 5725-1, 1994). Repeatability and robustness should not be confused. Robustness is the characterization of physical, chemical and biological variations (e.g. pH, temperature or time fluctuations, sample integrity, etc.) that can be tolerated without appreciably altering assay performance. Repeatability measurements essentially quantify and monitor robustness. Robust assays generally display a high degree of repeatability. Often referred to as a measure of precision, variation of binary results is in fact a combined measure of imprecision (random error) and inaccuracy (systematic error or bias) (Van den Bruel et al., 2007). This analytical estimation of repeatability is considered the baseline of the internal quality control program (IQC). To avoid any confusion later on, we further characterized this parameter as **analytical repeatability,** as opposed to field repeatability, described below. Repeatability of test

results is further refined during the field validation and continuously monitored during IQC.

The next analytical criteria to be considered are **analytic sensitivity** (ASe) and **analytical specificity** (ASp). ASe is defined as the ability of the test to detect and distinguish a minimum concentration of analyte from the sample matrix with a specified probability (OIE, 2009b). Also referred to as the limit of detection (LOD), ASe is estimated by serially diluting a solution of analyte in an analyte-free matrix of the same constitution as samples from the targeted population. Depending on the method, the OIE requires that the LOD be expressed as a number of organisms, genomic copies, colony-forming units, complement-forming units, or plaque-forming units for a defined sample volume or weight (i.e. concentration). Often the LOD is reported as a numerical quantity and not as a concentration (e.g. number of genome copies), but this practice is not recommended because an absolute number has no practicality since it ignores the critical information of dilution and/or sampling fraction.

ASp is defined as the ability of the test to test negative for components that may be present in the sample matrix and that are closely related to the targeted analyte (OIE, 2009b). In practice, ASp is estimated by testing a range of exposed animals or a group of animals infected by closely related organisms. Satisfactory levels of cross-reaction depend on the intended purpose of the test and on the relative frequency of the competing organism in the targeted population.

If the analytical evaluation of the assay is considered satisfactory, it is possible to obtain a *provisional recognition of validity* (i.e. OIE recognized but not OIE certified). To obtain this partial certification, OIE requires additional preliminary estimations of

diagnostic sensitivity, diagnostic specificity and reproducibility (see sections 1.2.1.2.2 and 1.2.1.2.3) on a small scale, and a direct comparison between the candidate (or index) test and a recognized standard test method. If satisfactory, the assay obtains the provisional recognition of validity for use in emergency situations and as part of bilateral agreements. However, it is important to recognize that bench (or analytical) validation does not replace the field (diagnostic) validation which will be the focus of later discussion.

1.2.1.2.2 Stage 2- Diagnostic performance characteristics

To report the efficiency or discriminatory accuracy of a diagnostic test, OIE requires the use of **diagnostic sensitivity** (DSe) and **diagnostic specificity** (DSp), two parameters introduced by Yerushalmy (1947) that rely on the principle of conditional probability. DSe (or true positive rate) is defined as the proportion (probability) of D+ individuals that test positive. DSp (or true negative rate) is defined as the proportion (probability) of D- individuals that test negative.

Others parameters of test accuracy maybe used but have with some limitations. **The efficiency** (Ef), or overall accuracy, is the overall proportion of tested individuals that are correctly classified. However, this parameter depends on the prevalence (i.e. prevalence-weighted average of DSe and DSp) and can strongly vary when (i) DSe and DSp differ substantially from each other and (ii) the prevalence of the targeted population substantially deviates from 50% (Alberg et al., 2004).

In 1950, Youden introduced a new summary index of discrimination efficiency renamed later as the **Youden index** (J). It is the average of "successes" (difference in

proportions of correctly and incorrectly classified individuals across disease groups(D+ and D-)). When J is above 0, samples have more chance to be correctly classified and the test is, therefore, considered useful (positive discrimination).

Diagnostic odds ratio (DOR) is another measure of diagnostic accuracy that reflects the ratio of the odds of being correctly classified over the odds of being misclassified (Lijmer et al., 1999). The greater DOR is above 1, the more useful the test is. If DOR is < 1, tested individuals are more likely to be misclassified (negative discrimination) than if the test had not been performed.

The advantages of J and DOR are that they are independent of the prevalence of D+ and are expressed as a single measurement of diagnostic performance that can be used as an outcome for test comparisons or meta-analysis studies (Glas et al., 2003). Conversely, the main drawback of these two parameters is the impossibility to differentiate classification performance in true D+ versus D- groups. Since a proper evaluation of "fitness for purpose" necessitates separate test performance indicators (e.g. DSe & DSp), it follows that these criteria are not part of in the OIE validation requirements. Also, it is possible to subsequently calculate Ef, J and DOR from DSe/DSp and others parameters (see section 1.5.2.1).

For tests with continuous outcome results, OIE requires the determination of a threshold or cut-off value(s) to dichotomize or categorize (i.e. negative, positive and/or intermediate) results. According to the purpose of the test, the cut-off can be selected based on overall accuracy, disease status-specific efficacy, clinical impact, or cost impact.

1.2.1.2.3 Stage 3- Ruggedness, reproducibility and repeatability refinement

An OIE prescribed test is expected to be used in multiple laboratories. Therefore, it is required to estimate the **ruggedness** and **diagnostic reproducibility** of the test. Ruggedness is defined as the capacity of a test to resist expected variation in test conditions across laboratories (OIE, 2009b). It identifies critical user variables that impact the transferability of a technique including sampling and storage procedures (Crowther et al., 2006). For instance, freezing (-80 ºC) was shown to increase DSe for samples submitted for ISAV testing by RT-PCR (Nérette et al., 2005a). A method may not be considered as "rugged" if it is highly affected by sampling conditions (Crowther et al., 2006). Influencing factors are user-dependent and consequently difficult to control.

The diagnostic reproducibility is a defined as a measure of the variation in test results obtained with the same method on identical test items in different laboratories with different operators using different equipment (between-laboratory consistency) (IS0 5725-1, 1994). Like robustness and repeatability, ruggedness identifies critical factors affecting test result variation across or between laboratories whereas reproducibility quantifies this variation. Similar to repeatability, reproducibility combines the measure of imprecision (random error) and inaccuracy (systematic error) in diagnostic results across laboratories. OIE guidelines encourage using replicate samples in each participating laboratory to further measure the repeatability and estimate its variation across laboratories.

1.2.1.2.4 Stage 4- Program implementation

Ultimately, the usefulness and success of a diagnostic test is revealed by its inclusion within regional, national or international programs. In essence, this is a form of external validation. OIE guidelines emphasize the inclusion of information from tested populations (i.e. prevalence of infection) to account for the probability to sample a D+ individual from the population. For instance, if the actual prevalence of infection is very low, the probability of a false positive result may be higher than the probability of getting a true positive result. Therefore, it is recommended that additional confirmatory and verification procedures be incorporated to further investigate potentially false test results (e.g. traceable analyte in positive control to identify cross-contamination).

The deployment of a recognized test involves a wide range of applications and requires extensive evaluation of assay ruggedness and production of large quantities of reliable reference reagents. The test must demonstrate consistent operating characteristics, regardless of the conditions of utilization (e.g. field conditions) and reagents must be homogeneous and stable (e.g. identical control aliquots).
Once approved by the OIE's Biological Standards Commission, the assay is listed with other validated and certified methods for a specified purpose. However, to be designated as a **prescribed** or **alternate** test for trade, the method must be recognized as useful at the regional, national and international level.

Thereafter, OIE requires that the validation status must be maintained by continuously monitoring the assay operating characteristics (maintenance pathway).

1.2.1.3 Assay validation maintenance

Following the initial evaluation of the operating characteristics of the assay, it is necessary to continuously control and monitor the test performance.

First, an **internal quality control** (IQC) program must be implemented to detect potential changes in accuracy over time. Control charts are recommended to graphically monitor the repeatability of control samples. Particularly applicable for tests with continuous outcomes, control charts include daily detailed data (DDD) and summary data charts (SDC) (Crowther et al., 2006). Secondly, laboratories are encouraged to participate in **external quality control** (EQC) programs. Most EQC programs require annual or bi-annual *proficiency testing* or *ring tests*. This is accomplished using panels of highly defined samples representing a range of analyte concentrations that are test 'blind' by the participating laboratories. The continuous assessment of repeatability and reproducibility is designed to ensure the consistency across time of test performance within- and between-laboratories.

The expression of a disease and associated pathogens may vary with time and location. For instance, it is expected that a test would not perform the same with different strains and/or virulence of the pathogen. As new strains or genotypes of pathogens arise, laboratories are advised to re-evaluate the ASe and ASp of the assay to ensure that any changes to the pathogen have not compromised the diagnostic accuracy of the test. In some cases, the test may require modification to accommodate these changes.

During the lifespan of an assay, improvements to the technique, reagents and/or equipment may enhance detection efficiency and/or the cost-effectiveness of the assay. In

such cases, the OIE does not require a total re-validation of the method. However, the OIE does require a demonstration of **method comparability**. It is recommended that equivalency be evaluated by assessing the agreement between the original and modified protocols run side-by-side at least 20 times suing a panel of samples representing a full range of analytes and concentrations. Similarly for depleted reagents, an overlap of paired runs (one with the reagent nearing depletion and its replacement) should be conducted to evaluate and confirm similar test performance.

Since confidence levels around estimated validation criteria are expected to improve with more evidence and increased frequency of testing, the OIE encourages the submission of additional information on test performance as its use becomes more common. However, it may not be appropriate to combine information from different over time since the population characteristics may have changed.

After demonstrating "fitness for purpose", a recently developed test may not necessarily be listed in the OIE Manual. The assay must prove to be useful and successful in surveillance programs. Before it can be recognized internationally, the test must prove its usefulness at regional or national levels.

1.2.2 National Aquatic Animal Health Program in Canada[1]

1.2.2.1 Origin and objectives

Canada is a member of the OIE and the World Trade Organization (WTO), and therefore accepts the OIE standards for trade of aquatic and terrestrial animals. Most of Canada's

[1]This section was adapted directly from the information available on the Canadian Food Inspection Agency (CFIA) and Department of Fisheries and Ocean Canada (DFO) websites (http://www.inspection.gc.ca/english/anima/aqua/aquaproge.shtml and http://www.dfo-mpo.gc.ca/aquaculture/health-sante-eng.htm#n1, respectively)

major trading partners have also adopted OIE international standards within their own aquatic animal health programs (e.g. National Aquatic Animal Health Plan in the US). For exportation, Canada may be required to attest that aquatic animals and their products are free of OIE reportable diseases. In reciprocal arrangements, Canada may require that imported aquatic products be controlled and certified to prevent the introduction of serious infectious diseases. Finally, numerous activities in Canada rely on aquatic resources (e.g. aquaculture, recreational fishing, ornamental hobbyists) and it is essential to limit disease spread within Canada to safeguard these industries. Therefore, Canada implemented its own regulatory framework to meet international standards, to protect Canadian aquatic resources (wild and farmed) from serious infectious diseases, and to maintain competitive access to international markets.

In 2005, the Government of Canada invested CDN$ 59 million over five years, with permanent funding thereafter, for CFIA and DFO to implement and deliver a National Aquatic Animal Health Program (NAAHP). This science-based program consists of regulatory measures needed to prevent, control and/or eradicate aquatic animal diseases of concern.

1.2.2.2 Implementation and organization

The implementation of the NAAHP is a joint responsibility of the Canadian Department of Agriculture and Agri-Food (responsible for the CFIA) and the Canadian Department of Fisheries and Oceans. CFIA was designated the lead federal agency for the NAAHP under the legislative authority of the *Health of Animals Act and Regulations*

(http://laws.justice.gc.ca/en/H-3.3?noCookie). The CFIA is responsible for disease surveillance protocols and control measures for reportable diseases, whereas DFO delivers and oversees the National Aquatic Animal Health Laboratory System (NAAHLS) and research to support NAAHP.

In Canada, the management of the wild fisheries and aquaculture is shared between the federal and provincial/territorial jurisdictions. The responsibilities and expertise within provinces/territories and industry provide complementary efforts which should minimize redundancy and gaps. To facilitate the development of the NAAHP, a National Aquatic Animal Health Steering Committee was created and includes representatives from all Canadian provinces and territories, Aboriginal and First Nations Peoples' associations, wild and farmed industry associations, veterinary associations, and academia.

The structure of the NAAHP is articulated around 8 interdependent actions:

1) **legislative and regulatory framework**: currently under the *Fisheries Act (Fish Health Protection & Fishery Regulations)*, the protection of wild and farmed fish against serious infectious diseases will in the future, be included under the *Health of Animal Act*, as for terrestrial animals;
2) **export health certification**: under CFIA responsibilities, the Agency will be releasing health certificates that conform to importing country requirements and/or OIE international standards;
3) **import control**: under CFIA responsibilities, the Agency will target control of imported products based on the risk of introduction that they represent (e.g. health

status of the exporting country, ultimate use in Canada), to prevent the introduction of reportable diseases;

4) **domestic disease control**: under CFIA responsibilities, the Agency will develop a framework, based on risk assessment and surveillance reports, to control the spread of serious infectious diseases within Canada and to implement emergency actions after introduction of a listed disease in Canada or part of Canada;
5) **surveillance**: under CFIA responsibilities, the Agency will coordinate surveillance programs to document freedom from notifiable diseases, to meet certification requirements, to assist the investigation of disease outbreaks, and to monitor the efficacy of control measures;
6) **risk assessment**: under CFIA responsibilities, the Agency will use available scientific information to identify the critical domains where regulatory control and policy development are needed;
7) **laboratory testing**: under DFO responsibilities, DFO will set and manage four national reference laboratories spread across the country (Gulf Fisheries Centre, Moncton, NB; Freshwater Institute, Winnipeg, MB; Pacific Biological Station, Nanaimo, BC; and Charlottetown Aquatic Animal Pathogen & Biocontainment Laboratory, Charlottetown, PEI) to develop and provide quality detection methods for listed diseases, confirmatory testing for suspicious findings by other laboratories, suspicious die-offs of wild or farmed fish, and surveillance programs to report freedom from OIE listed diseases;
8) **research**: under DFO responsibilities, DFO will try to address the knowledge gaps that impact regulatory decision making (e.g. risk assessment, diagnostic

validation). The research program is coordinated by the Centre of Expertise for Aquatic Animal Health Research and Diagnostics (Moncton, NB).

The use of diagnostic tests is essential in several of these interacting activities, including export certification, import control, surveillance, risk assessment, disease investigation, and emergency response. Therefore, it is required that any diagnostic method used to detect reportable diseases in the NAAHLS be validated according to the international guidelines and standards.

1.2.3 Applications of diagnostic test criteria

The OIE test validation guidelines describe 13 qualitative and quantitative assay validation criteria to assess the *fitness for purpose* of a test. Although post-estimation utilisation of these factors is not clearly stated in the OIE guidelines, these criteria have numerous applications in the interpretation of test results and to the decision-making process. Appropriate interpretation of a test result relies on the discriminatory accuracy of the test and on the conditions in which the test was applied (Shapiro, 1999). These conditions will include clear definitions related to the testing objectives, sampling design and unit of sampling (e.g. number of specimens or individuals sampled), unit of testing (e.g. pooled or multiple testing), unit of interpretation (e.g. individual, herd, region), and statistical method employed (e.g. statistical parameter, software) (Greiner & Gardner, 2000a). We will review the most common applications of diagnostic criteria, starting with the only one suggested by OIE: predictive value of a test result.

1.2.3.1 Predictive values of a test result

By definition, DSe and DSp are probabilities of a test result conditional on the fact that the health status is known. Practically, these parameters have limited use since the health status of any given animal prior to testing is rarely, if ever, known. Therefore, with the test result as the primary piece of information (i.e. the condition), the relevant probability of concern is: what is the probability of the true health status when the test result is known? This probability was introduced by Vecchio (1966) as the **predictive value** (PV) of a test result. The test result can be positive (PV+) or negative (PV-), and the status of interest can be either D+ (positive predictive value, PPV) or D- (negative predictive value, NPV). According to the test result and the status of interest, four conditional probabilities can be estimated: PPV+, NPV+, PPV-, and NPV-. The two parameters of most common interest are PPV+ and NPV-, conveniently called PPV and NPV. PPV is defined as the proportion or probability of a test positive individual to be D+, whereas NPV is the proportion or probability of test negative individual to be D-. In addition, predictive values can be estimated for different targeted populations using the test DSe, DSp (both assumed constant across populations) and expected prevalence (Pr). Deduced from the Bayes' theorem, it is possible to calculate predictive values as (Dohoo et al., 2009):

$$PPV = Pr\ DSe / (Pr\ DSe + (1-Pr)\ (1-DSp)) \quad (1)$$

$$NPV = (1-Pr)\ DSp / (Pr\ (1-DSe) + (1-Pr)\ DSp) \quad (2)$$

These conditional probabilities can also be estimated directly from a 2X2 table including the true status information (see section 1.5.2.1) as: PPV = a/(a+c) and NPV = d/ (b+d) (Akobeng, 2007a). If the prevalence of D+ is < 50%, DSe has more influence on NPV while DSp has more influence on PPV. In addition, predictive values are population parameters that strongly depend on prevalence. However, prevalence of D+ in populations is rarely known exactly before testing. It is advisable to explore different realistic levels of prevalence and assess the practical consequences based on the estimated predictive values. For instance, one of the most delicate scenarios for decision-makers is to have one positive test result during the surveillance of a truly D- population. With an almost perfect test (DSe = 99.99% and DSp = 99.99%) and a prevalence expected close to zero, the PPV will be very low. With a DSp of 99.99%, we expect one false positive result every 10,000 samples tested from a disease-free population. According to the sampling strategy, it would be biologically unlikely to have a single true positive individual to an infectious disease that has a tendency to spread within the population. Most likely to result from a cross-contamination (Wilson, 1997), it is recommended to set subsequent confirmation and verification procedures within the internal quality control strategy to investigate and track potential contamination and improve the level of confidence of the test result. No decision-threshold is set for predictive values since any decision process for animal health combines biological evidence and social, economic, and political considerations.

Although DSe and DSp are often believed to be constant, it is reasonable to believe that DSe and DSp vary with prevalence (Greiner & Gardner, 2000b). Modelling

analysis of predictive values has shown that the dual dependence of predictive values on prevalence (direct and indirect through DSe and DSp) tends to buffer their overall variation across prevalences (Brenner & Gefeller, 1997). Several software programs exist that facilitate the computation of predictive values. One of the most popular in veterinary applications is included in the epidemiological analysis package *Survey Toolbox* (Cameron & Baldock, 1998a &b, http://www.ausvet.com.au/content.php?page=res_software). A parallel version, *Bayes FreeCalc1*, was developed under the *BDFree* software to use posterior distribution parameters when DSe and DSp were estimated in a Bayesian framework (Johnson et al., 2004; http://www.epi.ucdavis.edu/diagnostictests/module02.html).

1.2.3.2 Likelihood ratio of a test result

Likelihood ratios (LR) are clinically relevant parameters that apply the information from DSe/DSp estimates. The likelihood ratio of a test result is defined as the proportion or probability of this test result among D+ individuals divided by the proportion or probability of this same test result among D- individuals (Akobeng, 2007b). Test result can refer to a defined category of measures when the test outcome is continuous or ordinal (category specific LR) or to a dichotomized test outcome (cutpoint LR). In this section, we restricted the discussion to cuptoint LR adapted to binary outcome test. Since a test results may be positive or negative, the LR of a positive test result (LR+) and the LR of negative test result (LR-) can be determined. LR+ provides information about how much more likely D+ individuals are to test positive compared to

a D- individual. Likewise, LR^- provides information about how much less likely D+ individuals are to test negative compared to a D- individual. In general, when LR is greater than 1, D+ individuals are more likely to get the test result than D-. When LR is less than 1, D- individuals are more likely to get the test result than D+. If there is insufficient evidence of LR being different from 1, the test does not perform differently in D+ and in D- individuals. Ultimately, the larger LR+ (above 1) and the closer LR- is to 0, the greater is the discriminatory power of the test.

For each cut-off value there are different DSe and DSp, so for each cut-off, LRs can be computed (Dohoo et al., 2009):

$$LR+ = DSe / (1-DSp) \quad (3)$$

$$LR- = (1-DSe) / DSp \quad (4)$$

LRs can also be computed directly from a 2X2contingency table (see section 1.5.2.1) as follows: $LR+ = a(c+d) / c(a+b)$; $LR^- = b(c+d) / d(a+b)$.

The main utilization of LR is for clinical interpretation. During the diagnostic process, a test method is employed by the clinician in combination with other information to *rule in* or *rule out* a suspected cause. Before the test is conducted, based on the animal history and clinical examination, and published information, the clinician should be able to have a rough estimate of the *pre-test probability* of the patient being D+. Based on the result category or continuous measure and associated test performances (i.e. DSe & DSp), the clinician can estimate the *post-test probability* of the patient being D+. Using the principles of the Bayes theorem, it is possible to calculate the **post-test odds** of the

probability to be D+ by multiplying LR with the **pre-test odds** of the probability to be D+:

$$Odds_{POST\text{-}TEST} = LR \times Odds_{PRE\text{-}TEST} \quad (5)$$

Based on the test result (i.e. positive or negative), LR^+ or LR^- are respectively used and the odds of the pre-test prevalence (*Pr*) is computed as $Pr / (1-Pr)$. Post-test probability can be back calculated from the post-test odds using $Odds_{POST\text{-}TEST} / (1 + Odds_{POST\text{-}TEST})$. However these calculations are intensive and not very intuitive, so a simpler method to determine the post-test prevalence is to use the nomogram developed by Fagan (Akobeng, 2007b). Fagan's nomogram is a graphical tool (Fig. 1.2) requiring two information entries (i.e. assumed pre-test prevalence and computed LR) to obtain directly the post-test prevalence based on the test results (Fagan, 1975). In general, for low prevalences (i.e. < 5%), the odds can be approximated to the prevalence itself (Dohoo et al., 2009).

This approach mainly targets continuous outcomes test with category specific LR but can be adapted for cutpoint LR of dichotomous test. The obtained estimations of post-test prevalences match then the estimations of predictive values. For instance, in surveillance or control programs settings, with no history or clinical information available, pre-test probability can be interpreted as the expected prevalence in the population when post-test probability is interpreted as the PPV based on the test result. According to the test result (i.e. positive or negative), the PPV of this test result can be assessed with the LR of the same test result (LR^+ or LR^-) with the assumed target

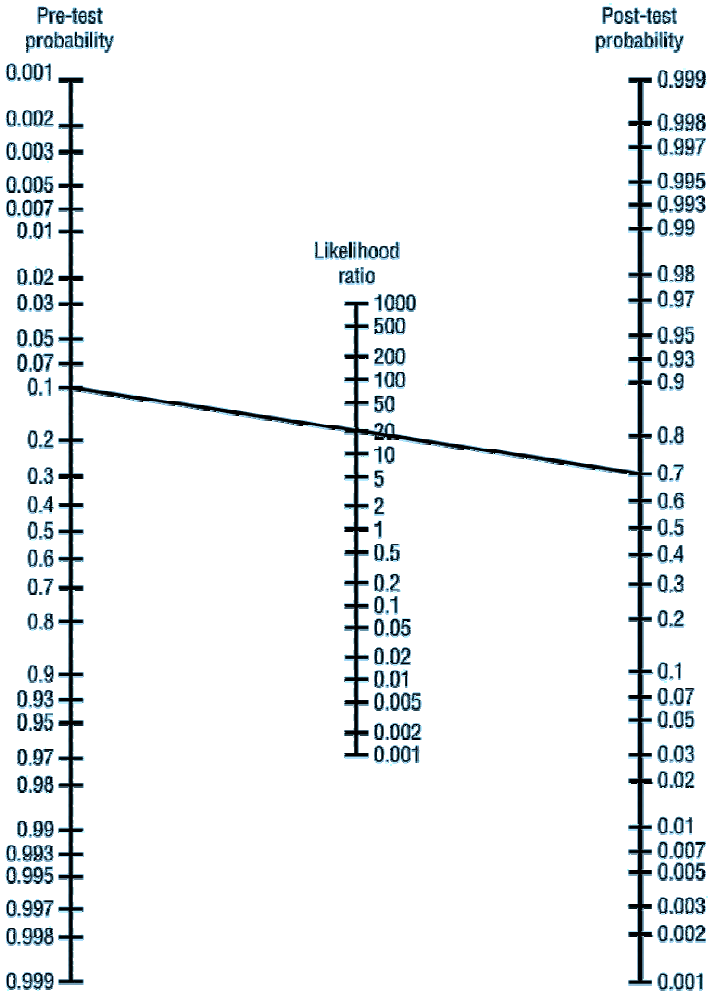


Fig. 1.2. The Fagan's nomogram (adapted from Deeks & Altman, 2004). By drawing a straight line between the assumed pre-test prevalence (or population prevalence), on the left axis, and the calculated likelihood ratio (LR) of the observed test result (negative or positive: LR^- & LR^+, respectively), on the middle axis, the post-test prevalence (or positive predictive value, PPV) is estimated at the intersection of this line with the right axis. For value of LR+ and LR- according to customary values of DSe & DSp refer to Appendices 1 & 2, respectively. *LR^+ is computed as DSe / (1-DSp), and LR^- as (1-DSe) / DSp.*

population prevalence (i.e. pre-test prevalence). This estimation, as for individual clinical patient post-test prevalence, can be done mathematically or graphically with Fagan's nomogram (Fig. 1.2). To facilitate LR calculation, separate estimation tables were constructed to assess LR^+ and LR^- as functions of DSe and DSp (Appendices 1 & 2, respectively).

1.2.3.3 Efficiency, Youden index, diagnostic odds ratio

The overall accuracy of a test is conventionally expressed using three indices: efficiency (Ef), Youden index (J) and diagnostic odds ratio (DOR). When the true status of the sampled individual is known, these indices can be computed directly from the 2X2 table (see section 1.5.2.1). As information about the true status is rarely known before hand, estimation of overall accuracy can be accomplished using DSe and DSp estimates. In this instance, Ef expresses the proportion of correctly classified individual (among D+ and D- combined) and is computed as a function of the DSe/DSp and prevalence (Pr) (Shapiro, 1999):

$$Ef = Pr\ DSe + (1-Pr)\ Dsp \qquad (6)$$

Ef is directly dependent on prevalence (i.e. prevalence-weighted average of DSe & DSp) and therefore has limited generalization compared to the status-specific accuracy measures of DSe and DSp (Alberg et al., 2004).

Next, J can be directly computed from DSe and DSp independent of the prevalence:

$$J = DSe + DSp - 1 \quad (7)$$

By restraining the added status-specific proportions of correctly classified individuals to the interval [-1,1] this index is a useful single measure of test accuracy to compare test methods, conduct meta-analysis on diagnostic performance studies, or select cut-off values for continuous tests.

Finally, DOR is also independent of prevalence and can be expressed as a function of DSe/DSp or PPV+/NPV- or LR+/LR-:

$$DOR = DSe\ DSp\ /\ (1-DSe)\ (1-DSp) = PPV+\ NPV-\ /\ (1-PPV+)\ (1-NPV-) = LR+\ /\ LR- \quad (8)$$

This index mirrors the ratio of correctly classified individuals over misclassified individuals, thus providing a measure of the overall discrimination performance of the test. DOR is a symmetrical parameter and therefore allows for conditional interpretations, on either the health status or the test result. As for J, DOR is a convenient single indicator of diagnostic accuracy that facilitates test comparisons, meta-analyses, and selection of a threshold (cut-off) (Glas et al., 2003). In addition, DOR is particularly meaningful when interpreting variation in test performance across studies using logistic regression models (Glas et al., 2003).

1.2.3.4 Sample size calculation for demonstration of freedom of disease

A major assumption for a sample size calculation to demonstrate freedom from infection/disease is that the diagnostic test used to screen individuals is perfect (100% DSe & DSp) (Cannon & Roe, 1982). However, test methods are rarely perfect and sample size calculations need to be adjusted for misclassification bias. In this instance, a sample size calculation for freedom of disease can be corrected by dividing the initially computed sample size (*n*) by the DSe:

$$n' = n / DSe = (\ln \alpha / \ln (1-Pr^*)) / DSe \qquad (9)$$

where α refers to the type I error (complement of confidence level) and Pr^* refers to the minimum expected prevalence if the disease is present in the population. Potential misclassification of D+ individuals is adjusted by increasing the sample size to ensure that, if present in the sample set, at least one D+ will be correctly classified.

Open access software packages are available to compute sample sizes to demonstrate freedom of disease in frequentist or Bayesian frameworks respectively: *FreeCalc, Survey Toolbox* (Cameron & Baldock, 1998a&b); *Bayes FreeCalc2*, *BDFree* (Johnson et al., 2004).

1.2.3.5 True prevalence estimation

Another useful application of DSe and DSp is to calculate the **true prevalence** (TP, proportion of D+ individuals) by adjusting the **apparent prevalence** (AP, proportion of positives) generated by the test. The AP can be expressed as a function of DSe and DSp using the sum of the proportion of true positives and the proportion of false positives in the population of prevalence TP:

$$AP = TP\ DSe + (1-TP)\ DSp \qquad (10)$$

From Eq. (2), it is possible to arithmetically deduce TP as a function of the three other factors (DSe, DSp, AP) (Rogan & Gladen, 1978):

$$TP = (AP + DSp - 1) / (DSe + DSp - 1) \qquad (11)$$

where DSe +DSp is expected to be greater than 1.

Open access software available in the *Survey Toolbox* package (Cameron & Baldock, 1998a&b) can be used for these calculations.

1.2.3.6 Determination of fitness for purpose of a testing strategy

Ultimately, the user aims to maximize the level of certainty (i.e. predictive value) to facilitate the decision-making process according to the intended purpose. Predictive

values are influenced by population factors (i.e. prevalence) over which the test user has very limited control. However, according to the testing purpose, the sampling strategy can be optimized (e.g. targeted sampling strategy) to increase prevalence for instance. The targeted operating characteristics of a testing strategy of a test may vary substantially and fit different purposes according to the intrinsic performance of a test, potential test combinations and interpretations (e.g. parallel or series), and the unit of interest (e.g. individual, herd).

Commonly, the operating characteristics of interest are selected based on the mnemonic "SnNout and SpPin" (Akobeng, 2007a). For initial screening purposes, a well suited test is expected to be highly **sen**sitive to increase the confidence of a **n**egative test result (NPV^-) and rule **out** disease (**SnNout**). For confirmation purposes, a well suited test is expected to be highly **sp**ecific to increase the confidence of a **p**ositive test result (PPV^+) and rule **in** disease (**SpPin**).

According to OIE, 6 different categories of intended purposes exist and require different discriminative abilities (OIE, 2009b). Table 1.3 summarizes these purposes and the required discriminative skills to fit them.

1.2.3.7 Interpretation of diagnostic parameters at the herd level

Herd interpretation - Conventionally in veterinary epidemiology, the term *herd* refers to a group or aggregate of animals (e.g. tank or netpen group of fish). If the unit of sampling and testing is at the herd level (unit of concern), the interpretation of a test result, follows the same approach (e.g. a count of sealice larvae in a water sample from a

Table 1.1
Objectives of the minimization and/or maximization of test parameters according to the six OIE diagnostic intended purposes.
Although, this table is presented at the individual level, it can also be interpreted at the herd level.

Intended purposes	Objective	Target	Test strategy criteria requirement
1. Demonstration of population 'freedom' from infection	Min. FP	Max. PPV^+ & LR^+	Maximize DSp, r/R^{D-}
2. Demonstration of freedom from agent in trade product	Min. FN	Max. NPV^- & Min. LR^-	Maximize DSe, r/R^{D+}
3. Eradication of infection from defined populations	Min. FP (economic concern) Min. FN (zoonotic concern)	Max. PPV^+ & LR^+ Max. NPV^- & Min. LR^-	Maximize DSp, r/R^{D-} Maximize DSe, r/R^{D+}
4. Confirmatory diagnosis of clinical cases	Min. FP	Max. PPV^+ & LR^+	Maximize DSp, r/R^{D-}
5. Estimation of prevalence of infection for risk analysis	Min. FP & FN (estimation error)	Max. PPV^+,NPV^-, LR^+ Min. LR^-	Maximize DSe/DSp, r/R
6. Determination of immune status (post-vaccination)	Min. FP & FN (estimation error)	Max. PPV^+,NPV^-, LR^+ Min. LR^-	Maximize DSe/DSp, r/R

FP: fraction of false positive
FN: fraction of false negative
PPV^+: positive predictive value of a positive test, probability to be diseased if test positive
NPV: negative predictive value of a negative test, probability to be non-diseased if test negative
DSe: diagnostic sensitivity, probability to test positive if diseased
DSp: Diagnostic specificity, probability to test negative if non-diseased
r: repeatability, test results agreement within a laboratory for a non-diseased (D-) or diseased individual (D+)
R: repeatability, test results agreement between laboratories for a non-diseased (D-) or diseased individual (D+)
LR^+: likelihood ratio of a positive test, ratio of the probability to test positive if diseased and the probability to test positive if non-diseased
LR^-: likelihood ratio of a negative test, ratio of the probability to test negative if diseased and the probability to test negative if non-disease

salmon cage is interpreted at the cage level), as previously described (Dohoo et al., 2009). However, if the unit of sampling and testing differ from the unit of concern (e.g. individual salmon tested but level of concern is the cage), the herd level interpretation of the test performance relies on DSe and DSp, and also on the prevalence of D+ within the herd (Pr), the number of animals sampled (n), and the cut-off number of reactors (k) (i.e. animals testing positive) needed to declare the herd infected.

For individual fish from an infected cage, the probability to test positive is expressed by the apparent prevalence of the infected cage (AP^+):

$$AP^+ = Pr\ DSe + (1-Pr)\ (1-DSp) \qquad (12)$$

Consequently, the probability for a fish to test negative in this infected cage is the complement probability: $1- AP^+$. Assuming that sampled fish are independent (i.e. no clustering among samples) and that the lack of replacement of sampled fish does not substantially impact test result probabilities (i.e. there is a small sample size compared to total population size), the probability that the *n* sampled fish will test negative is: $(1- AP^+)^n$.

If a single reactor is required to call the cage positive (k = 1), then the probability to declare the cage positive given it is infected (**herd sensitivity**, HSe) is the complement probability of all sampled fish testing negative:

$$HSe = 1 - (1 - AP^+)^n \qquad (13)$$

When the number of required reactors is ≥ 1, HSe is computed using the binomial probability distribution (Martin et al., 1992) assuming that the individual probabilities of a test result are not impacted by the lack of replacement during sampling (binomial approximation) and the health status of sampled fish are conditionally independent:

$$HSe = 1 - \Sigma_{0}^{k-1} C_{k-1}^{n} (AP^{+})^{k-1} (1 - AP^{+})^{n-(k-1)} \quad (14)$$

Note: C_b^a *is computed as a! / (a-b)! b!; and a!= a * a-1 * ... * 1*

The binomial probability distribution approximation is only valid if the sample size (n) is lower than 20% of the total population size (N) and sampling with replacement can be extrapolated (i.e. probability of a fish to be samples is constant) (Christensen & Gardner, 2000). When the sample size is larger, this approximation is not acceptable and HSe and HSp have to be computed using a hypergeometric probability distribution approach (Martin et al., 1992).

For individual fish from a non-infected cage, the probability to test negative is simply computed as the complement of the apparent prevalence in this population when Pr = 0% using Eq. (12): $Prob(T^{-}) = 1 - AP^{+} = DSp$.

Assuming that D- fish are independent, that the binomial approximation is valid and a single reactor is required (k =1) for a cage to be classified as test-positive, the probability to declare the cage negative (i.e. *n* samples tested negative), given it is free from infection (**herd specificity**, HSp) is:

$$HSp = DSp^{n} \quad (15)$$

When the number of required reactors is ≥ 1, HSp is computed as (binomial approximation):

$$HSp = \Sigma_{0}^{k-1} C_{n}^{k-1} (1 - DSp)^{k-1} (DSp)^{n-(k-1)} \quad (16)$$

If the binomial approximation is not acceptable, the hypergeometric probability approach can again be used (Martin et al., 1992). In addition, it is expected that, conditional on the health status, sampled fish are dependent (clustered) (Christensen & Gardner, 2000). Adjustment for clustering are beyond the scope of this discussion and are covered elsewhere (Donald et al., 1994).

According to the individual test accuracy (DSe & DSp), the population characteristics (Pr), and surveillance design (n & k), HSe and HSp may vary substantially (Christensen & Gardner, 2000). The influence of each factor is described in Table 1.2. Developed by Jordan in 1996 (http://www.vetschools.co.uk/EpiVetNet/Sampling_software.htm), an open access software called *HERDACC* version 3 enables the computation of HSe and HSp with binomial or hypergeometric probability distribution approaches for different values of within herd prevalence (Pr), sample size (n) and number of reactors (k). A major advantage of this software is the possibility to compute the number of reactors necessary for specified HSe or HSp (Jordan, 1996). A more flexible version of HERDACC was developed in a Bayesian framework that takes lack of independence of test results into

Table 1.2
Influence of diagnostic, population and surveillance parameters on the herd sensitivity and specificity (HSe & HSp) (adapted from Christensen & Gardner, 2000).

	Change	HSe	HSp	Comments
Diagnostic sensitivity (DSe)	up	up	no influence	
	down	down	no influence	
Diagnostic specificity (DSp)	up	down	up[a]	[a]for n fixed
	down	up	down[a]	[a]for n fixed
Number of animal tested (n)	up	up[b]	down[c]	[b]minimal if Pr>.4; [c]for DSp < 1
	down	down[b]	up[c]	[b]minimal if Pr>.4; [c]for DSp < 1
Number of reactor threshold (k)	up	down	up	
	down	up	down	
Within herd prevalence (Pr)	up	up	no influence	For n fixed & DSe>(1-DSp)
	down	down	no influence	For n fixed & DSe>(1-DSp)
Diseased correlation (cov^+)	up	down	no influence	
	down	up	no influence	
Non-diseased correlation (cov^-)	up	down[d]		[d]minimal if k>2
	down	up[d]		[d]minimal if k>2

account (Jordan & MacEwen, 1998). Donald et al. (1994) also developed a software (*AGG.exe*) that explores effects of correlation using a betabinomial probability distribution. Finally, *Freecalc* (from *Survey Toolbox*) uses the maximum number of positive test results to declare a herd free of disease as the required number of reactors (Cameron & Baldock, 1998a&b).

It is also possible to compute herd predictive values (HPPV & HNPV) (Martin et al., 1992), likelihood ratios (HLR), efficiency (HEf), Youden index (HJ), diagnostic odds ratio (HDOR) and herd true prevalence (HTP) by replacing the individual level parameters (DSe, DSp, and Pr) with HSe, HSp and the prevalence of infected herds (HPr) (Christensen & Gardner, 2000).

1.2.3.8 Interpretation of diagnostic parameters with pooled samples

One strategy to increase the proportion of total animals screened and/or reduce the testing costs per individual animal is to pool the samples (e.g. tissues or fluids) from several individuals and test them as a single unit. This practice is particularly attractive during surveillance programs when the probability of sampling a D+ fish is low (i.e. low prevalence). In this instance, the unit of sampling (fish) is now different from the unit of testing (i.e. pool of samples) leading to an interpretation at the pool level or at the herd level (several pools tested in a herd).

At the pool level, it is possible to use individual estimates of the test DSe and DSp to predict the DSe and the DSp of a pool of *m* randomly collected samples (PSe and PSp, respectively). These calculations account for the probability to sample D+ individuals

from the herd. The following calculations do not apply for pools of samples that include specimens of different origin or with different probabilities to be D+. Assuming that no clustering occurred within pools (conditional independence among sampled fish), the probability that a pool of *m* D- fish test negative (PSp) from a free population (Pr =0) is (Christensen & Gardner, 2000):

$$PSp = DSp^{m} \qquad (17)$$

This formula assumes that the binomial approximation is acceptable (i.e. fish are sampled without replacement and the sample size is < 20% total population size). Computation of the pool sensitivity (PSe) can be very complex due to its dependence on numerous factors and can end very complex (Muñoz-Zanzi et al., 2006). We suggest here one approach that requires to be validated. Using the assumptions of no fish clustering, binomial approximation, homogeneous pools of samples, and the power of individual detection is not influenced by dilution (i.e. a single true positive in the pool is sufficient to yield a positive result for the pool), the probability that a pool of *m* fish with at least one D+ ($k \geq 1$) that tests positive (PSe) follows:

$$PSe = 1 - \sum_{k=1}^{m} C_k^m \, (Pr\,(1-DSe))^k \, ((1-Pr)\,DSp)^{m-k} \qquad (18)$$

The assumption of no dilution effect is frequently incorrect. Depending on the pooling technique, the final pooled specimen can be either diluted (e.g. mixture of specific volumes of biological tissues or fluids) or concentrated (e.g. sequential dipping of

sampling swabs in the same set volume of transport media). Since it is associated with the analytical sensitivity (ASe) and potential inhibitor concentration, PSe decreases with diluted pools (especially when prevalence is low) (Christensen & Gardner, 2000), while DSe may increase with concentrated pools. Conversely, concentrated pools should increase the concentration of a potential contaminant and therefore decrease PSp (Christensen & Gardner, 2000), while diluted pools may increase PSp.

The dilution or concentration effect of pooling generates some clustering effect among samples according to their biological characteristics (Dohoo et al., 2009). For instance, the detection of D+ samples may depend on the concentration of analyte, the non-detection of D- samples may depend on the degree of exposure to different contaminants, and the interaction between D- and D+ samples may change the detection characteristics, depending on the homogeneity of mixing and increased complexity of the sampling matrix. Therefore, the assumption that test performance is independent among samples (conditional or not on the health status) associated with the dilution/concentration effect and the homogeneity of the pool is rarely legitimate.

Furthermore, by increasing the number of animals screened for a similar cost of analysis, sample pooling can be used to increase the HSe (and decrease of HSp) (Christensen & Gardner, 2000). In this instance, the unit of sampling (e.g. fish) is different from the unit of testing (pool), and different from the unit of interpretation (herd level). By binomial approximation, the probability that a pool of *m* samples tests negative when the herd prevalence is Pr, the complement of the pool apparent prevalence (PAP) is calculated as follows:

$$1\text{-}PAP = (1\text{-}(1\text{-}Pr)^{m})\ (1\text{-}PSe) + (1\text{-}Pr)^{m}\ PSp \qquad (19)$$

The herd pool sensitivity (HPSe) is then calculated including the number of pools tested (r) and the number of required reactor pools (k):

$$HPSe = 1\text{-}\ \Sigma_{0}^{k-1}\ C_{k-1}^{r}(1\text{-}(1\text{-}PAP))^{k-1}(1\text{-}PAP)^{r-(k-1)} \qquad (20)$$

Alternatively, the herd pooled specificity is calculated as:

$$HPSp = 1\text{-}\ \Sigma_{0}^{k-1}\ C_{k-1}^{r}\ (1\ -\ PSp)^{k-1}(PSp)^{r-(k-1)} \qquad (21)$$

For the particular case where k = 1, the HPSe and HPSp are, respectively (Christensen & Gardner, 2000):

$$HPSe = 1\text{-}\ [(1\text{-}(1\text{-}Pr)^{m})\ (1\text{-}PSe) + (1\text{-}Pr)^{m}\ PSp]^{r} \qquad (22)$$

$$HPSp = PSp^{r} \qquad (23)$$

A program to compute within herd prevalence from pool testing results (at least two sample pools tested) was developed for both frequentist and Bayesian approaches and is reviewed elsewhere (Cowling et al., 1999).

1.2.3.9 Evaluation of multiple testing performances

Utilization of multiple assays may improve the overall performance of testing and may, in some instances, reduce operating cost. More than two test methods can be used, however, the interpretation and the estimation of the combined operating characteristics may be complicated. The following discussion will be limited to the combination of 2 tests for the sake of brevity.

Paired results from 2 tests can be interpreted either **in parallel** or **in series**. For parallel interpretation of paired results, an individual is deemed positive if it yields at least one positive result; and a sample is deemed negative if it yields two negative results. For series interpretation of paired results, an individual is deemed positive if it yields two positive results; and a sample is deemed negative if it yields at least one negative result.

The proportions of the four possible test result combinations (i.e. ++, +-, -+, --) can be expressed using individual DSe and DSp of the respective test and a covariance factor representing the test dependence, conditional on the health status (cov+ & cov-, respectively) (Table 1.3). Therefore, parallel and series DSe/DSp (DSe_P/DSp_P & DSe_S/DSp_S, respectively) are as follows:

$$DSe_P = 1- [(1-DSe_1)*(1-DSe_2) + cov+] \quad (24)$$

$$DSp_P = DSp1 * DSp2 + cov- \quad (25)$$

$$DSe_S = DSe1 * DSe2 + cov+ \quad (26)$$

$$DSp_S = 1- [(1-DSp_1)*(1-DSp_2) + cov-] \quad (27)$$

Table 1.3
Compliance table between 2 assays where cells are expressed with diagnostic sensitivity (DSe) and a covariance factor (cov+) for infected/diseased; and diagnostic specificity (DSp) and covariance factor (cov-) for non-infected/non-diseased.

Infected/Diseased

		Test 2	
		Positive	*Negative*
Test 1	*Positive*	DSe_1*DSe_2+cov+	$DSe_1*(1-DSe_2)-cov+$
	Negative	$(1-DSe_1)*DSe_2-cov+$	$(1-DSse_1)*(1-DSe_2)+cov+$

Non-infected/Non-diseased

		Test 2	
		Positive	*Negative*
Test 1	*Positive*	DSp_1*DSp_2+cov-	$DSp_1*(1-DSp_2)-cov-$
	Negative	$(1-DSp_1)*DSp_2-cov-$	$(1-DSp_1)*(1-DSp_2)+cov-$

Table 1.4
Compliance table between 2 assays where cells are expressed with observed proportions of paired results for infected/diseased (P^{D+}) and non-infected/non-diseased (P^{D-}).

Infected/Diseased (D+)

		Test 2		
		Positive	*Negative*	
Test 1	*Positive*	P^{D+}_{++}	P^{D+}_{+-}	$P^{D+}_{+.}$
	Negative	P^{D+}_{-+}	P^{D+}_{--}	$P^{D+}_{-.}$
		$P^{D+}_{.+}$	$P^{D+}_{.-}$	*1*

Non-infected/Non-diseased (D-)

		Test 2		
		Positive	*Negative*	
Test 1	*Positive*	P^{D-}_{++}	P^{D-}_{+-}	$P^{D-}_{+.}$
	Negative	P^{D-}_{-+}	P^{D-}_{--}	$P^{D-}_{-.}$
		$P^{D-}_{.+}$	$P^{D-}_{.-}$	*1*

When the two tests are **conditionally independent**, the covariate factors (cov+ & cov-) are nil. In other words, for a defined infection class (i.e. condition), the results of one test does not depend on the result of the other test (i.e. the proportion of a test result is the same in the two result categories of the other test) (Dohoo et al., 2009). Note that if DSe (or DSp) of one test is perfect (100%), the two tests are conditionally independent in D+ (D-).

If the tests are conditionally independent, parallel interpretation improves the DSe for screening purposes; and series interpretation improves the DSp for confirmatory purposes. When tests are conditionally dependent (cov+ and/or cov- not nil), combined performances may partially increase or decrease according to the degree of correlation between tests. For D+ animals, when tests are positively correlated (cov+ > 0), DSe_P decreases and DSe_S increases compared to when the tests are conditionally independent. Reciprocally, when tests are negatively correlated (cov+ < 0), DSe_P increases and DSe_S decreases. For D- animals, if tests are positively correlated (cov- > 0), DSp_P increases and DSp_S decreases; and if tests are negatively correlated (cov- < 0), DSp_P decreases and DSp_S increases. When the true status of the specimens is known, cov^+ and cov^- can be estimated by comparing the observed (Table 1.4) and the expected proportions of paired result combinations (Table 1.3) as follows:

$$\text{Cov+} = P^{D+}_{++} - (DSe_1 * DSe_2) \quad (28)$$

$$Cov\text{-} = P^{D-}_{--} - (DSp_1 * DSp_2) \quad (29)$$

where P^{D+}_{++} is the observed proportion of the two test results that are positive in the D+ group while P^{D-}_{--} is the observed proportion of the two test result that are negative in the D- group. When the true status is unknown, latent class modelling procedures can be used to estimate cov+ and cov- under specified conditions (see section 1.5.2.2.2).

Statistical procedures to test for test dependence (conditional or not) include: Cohen's kappa significance test (compared to 0), odds ratio significance test (compared to 1), logistic regression modelling (log odds ratio), loglinear/association/quasi-symmetry modelling.

Multiple tests can be used **simultaneously**, but **sequential** testing provides more practical cost reductions by processing the less expensive test first. For parallel interpretation, only samples initially testing negative are subsequently submitted to the second test. For series interpretation, only samples initially testing positive are subsequently submitted to the second test. Assessment of the most cost-effective strategy to combine and interpret multiple tests in a Bayesian framework is discussed in detail by Geisser & Johnson (1992).

1.2.3.10 Further assessments of test efficacy

Although not specified in OIE guidelines, alternative evaluation methods to assess test efficacy or usefulness can be used. For instance, the test can be evaluated on the basis of clinical patient outcomes or economic consequences of the test result (Van den Bruel et al., 2007). the health management of an animal is, however, often primarily directed by economic concerns especially in the international trade of animal food and products. Nérette et al. (2008a) conducted detailed analysis for selection of the most economic

strategies for ISAV surveillance programs in Atlantic salmon aquaculture. According to the intended objectives, the factors to consider when in designing and comparing testing strategies in animal production are:

1. *Animal production structure* – The organisation of the industry and the different sectors (in time and in space) of the production system must be clearly defined. For instance, Nérette et al. (2008a) identified four separate production sectors in salmon industry: seawater broodstock, freshwater broodstock, pre-smolt (freshwater) and grow-out salmon (saltwater).
2. *Estimation of test performance at the fish level* – For all test methods of interest, estimates of individual DSe and DSp must be known and their validity in the different production sectors identified (i.e. external validity). Performances of test combination (usually two) should be explored and estimated (see section 1.2.3.9). For instance, Nérette et al. (2008a) investigated parallel and series interpretation of paired tests.
3. *Definition of a case and level of interpretation* – According to the production sector, identification of the appropriate unit of concern (fish, tank, cage or hatchery level) and the corresponding criteria to classify the unit of concern as test-positive (e.g. number of reactors). For instance, Nérette et al. (2008a) interpreted broodstock at the individual level, while pre-smolt and grow-out salmon were interpreted at the pen level.
4. *Test criteria for comparison.* When the individual fish is the unit of concern appropriate for the intended purpose, the testing strategies are compared based on PPV^+ or NPV^- and their associated costs. When the herd is the unit of concern

appropriate for the intended purpose, the testing strategies are compared based on $HPPV^+$ or $HNPV^-$ and their associated costs. If the clinical patient outcome is the primary interest, the testing strategy would be based on the probability of a favourable patient outcome.

5. *Selection of testing strategies*- For individual interpretation, all fish must be sampled and tested. Therefore, the best testing strategy is based on the best detection performances and lowest costs across different levels of realistic infection prevalence. For herd interpretation, the best testing strategy is based on the lowest associated cost for targeted detection performance. Only a portion of fish are sampled and tested in a herd. Therefore, the optimal detection performance ($HPPV^+$ & $HNPV^-$) depends on the number of sampled fish, if the samples are pooled or not, the minimum required number of reactors, the within herd prevalence (Pr) and the herd level prevalence (HPr). Consequently, different sample, pool and reactor sizes should be considered and compared across different levels of Pr for minimum targeted values of HSe and HSp and specific levels of HPr. For instance, Nérette et al. (2008a) compared different sample (not pooled) and reactor sizes suited for a minimum HSe and HSp of 95% and the assumed HPr of 0.1%.

Evaluation of DSe and DSp is the initial step and critical for appropriate decisions regarding the use of diagnostic tests. Testing strategies cover a wide range of designs according to the units of sampling, testing, and interpretation, and the access to multiple

tests and concerns associated with the classification uncertainty (clinical and economic consequences).

1.3 Infectious Salmon Anaemia

1.3.1 Historic and modern challenges

Numerous review articles provide a detailed and extensive source of information regarding infectious salmon anaemia (ISA) (Cipriano, 2002; Mjaaland et al., 2002; Rimstad & Mjaaland, 2002; Kibenge et al., 2004; Martin, 2007; OIE, 2009b). A good indication of the importance of ISA is reflected by the frequency of peer-reviewed publications addressing this specific disease. Since the first known occurrence of the disease in 1984 in Norway (Thorud & Djupvik, 1988), ISA has been increasingly studied with sporadic publication peaks corresponding to the progression or new geographical occurrence of the disease (Fig. 1.3). A detailed review of the information relevant to diagnostic use and evaluation for ISA will be the focus of this section.

ISA is frequently a lethal disease of farmed Atlantic salmon, *Salmo salar* L., caused by an orthomyxovirus (Falk et al., 1997). A member of the genus *Isavirus*, it is related but distinct from *Influenza* viruses (Kawaoka et al., 2005). Infectious salmon anaemia virus (ISAV) is described as an enveloped, spherical particle of approximately 90-140 nm diameter (Hovland et al., 1994) and containing 8 genomic segments of single-stranded negative-sens RNA (Mjaaland et al., 1997). More conserved and providing higher amplification success, the 8th segment is targeted for amplification in nucleic acid

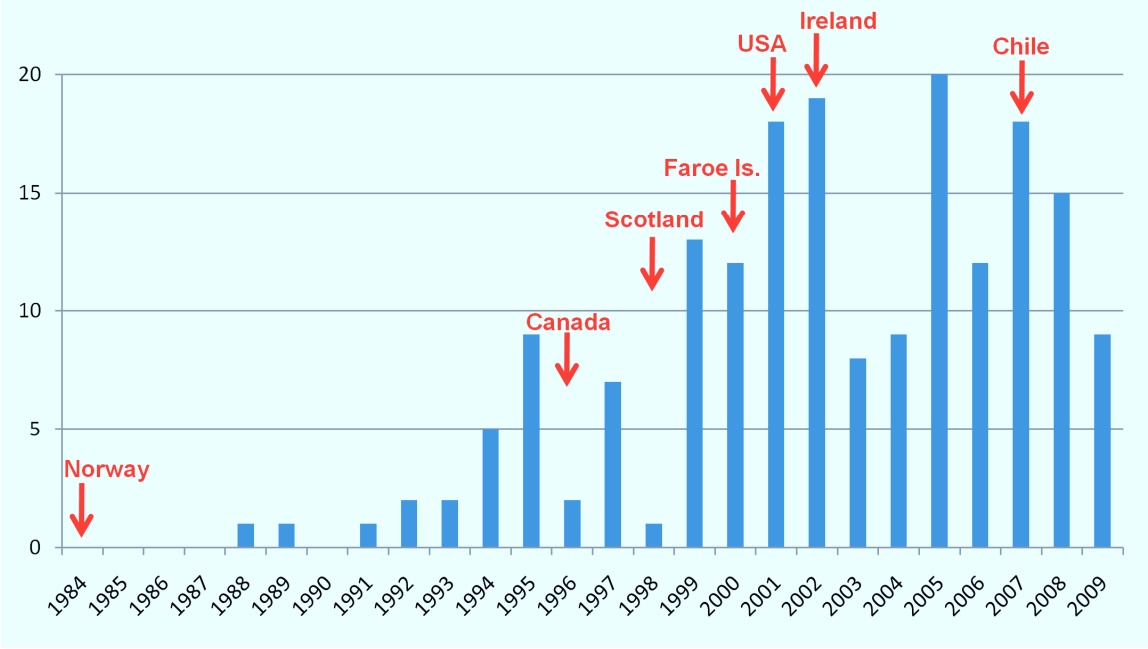


Fig. 1.3. Count per year of published references for infectious salmon anaemia (ISA). Red arrows indicate the year of first record of ISA according to the indicated country.

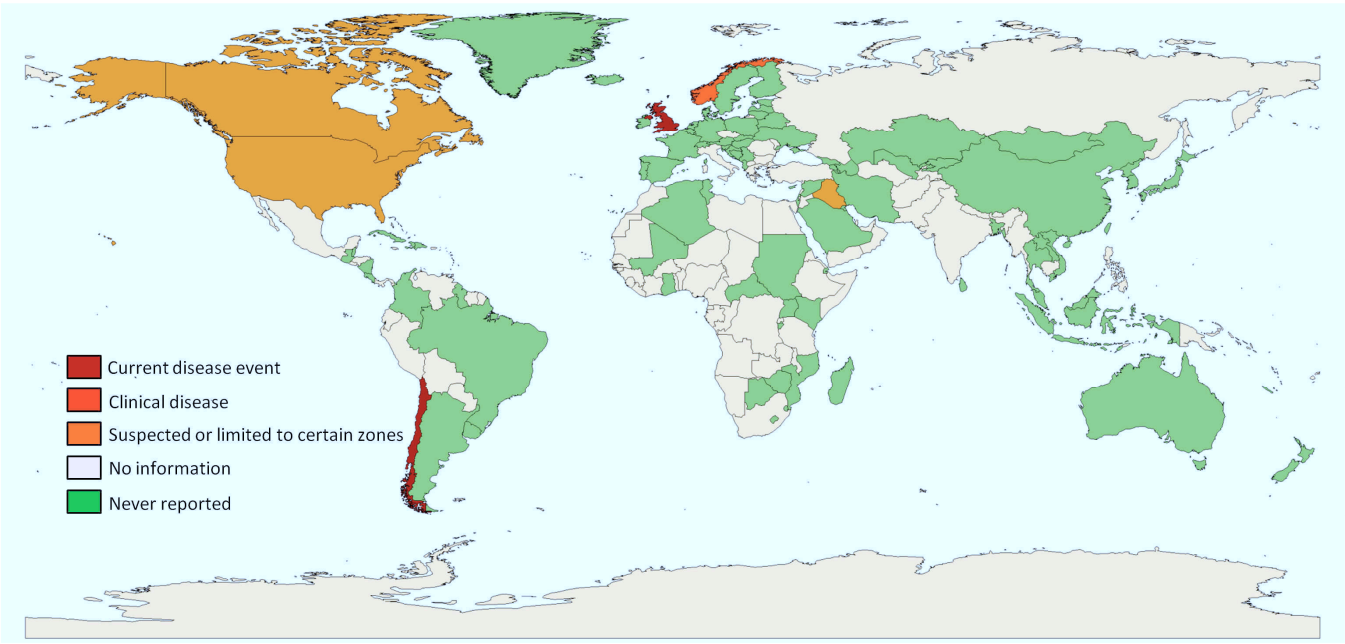


Fig. 1.4. Infectious salmon anaemia distribution map based on 2005-2009 reports.
Adapted from OIE, WAHID website:
http://www.oie.int/wahis/public.php?page=disease_status_map&disease_type=Aquatic&disease_id=160&sta_method=semesterly&selected_start_year=2008&selected_report_period=1&selected_start_month=1&page=disease_status_map.

detection (NAD) methods, while the 6th segment is used for virus typing by sequencing the highly polymorphic region (HPR) (Rimstad et al., 2001). Recently, the 5th segment is also used to increase the resolution of the genotyping (Nellie Gagné, pers. com.). The viral genome has been described as coding for at least 10 proteins (9 structural and 1 non-structural) (Kibenge et al., 2004). Among these proteins, two are surface glycoproteins which compose resembling mushroom-shaped projections that are of particular interest for diagnostic tests. A hemagglutinin-esterase (HE), coded by segment 6 (Rimstad et al., 2001), is responsible for receptor-binding (hemagglutinin) and receptor-destroying (esterase) activity, and a fusion protein (F), coded by segment 5 (Aspehaug et al., 2005), is putatively responsible for the fusion between the viral and cellular membranes.

The antigenic properties of the surface proteins induce a protective immune response in exposed hosts associated with the production of anti-ISAV specific antibodies (Falk & Dannevig, 1995). The presence of antibodies against ISAV suggests attractive applications for both direct and indirect detection methods (Falk et al., 1998). The physical and chemical properties of these proteins elucidate the virus pathophysiology and further explain the limited host range of the virus (i.e. salmonids) (Cipriano, 2002). For instance, the thermolability of the virus to higher temperatures (i.e. > 25°C; Falk et al., 1997) disproves the hypothetical infection of homeothermic vertebrates, including humans. With no zoonotic potential, ISAV mainly represents an economical threat for salmon farming due to mortalities, depopulation losses and surveillance costs. In 1999, the annual economic impact of ISA was estimated at US$ 14 million in Norway, US$ 11 million in Eastern Canada, and around US$ 15 million in Scotland (Hasting et al., 1999). Among the Atlantic salmon production areas, some

countries remain ISA-free (e.g. Australia and New-Zealand) and other are considered currently free (e.g. Faroe Islands) (Fig. 1.4). Given international trade concerns, specifically significant production losses and potential for introduction to free areas, ISA was listed as a concern by OIE in 1990 (Mjaaland et al., 2002) and became officially reportable in 2003 (Martin, 2007). In Canada, ISA has been present on the Atlantic coast since 1996 (Mullins et al., 1998), but the Pacific coast (British Colombia) is officially free of viral infection (WAHID OIE website). For this reason, the Canadian National Aquatic Animal Health Program (NAAHP) is responsible for maintaining barriers to the spread of ISA between the two coasts through coordinated surveillance programs by assessing and controlling practices contributing to the risk of transfer.

Control and intervention strategies are limited for a viral disease such as ISA. Scotland was apparently able to eradicate the disease using aggressive culling based on an intensive surveillance program (Stagg et al., 2001). Recently, however, Scotland experienced new clinical outbreaks of ISA in the Shetland Islands (WAHID OIE website). Development and evaluation of detection methods has been identified as critical for the efficient control of ISA (RSE, 2002).

1.3.2 Organ of choice for testing for ISA

Tissue collection is a key element to the successful detection of a pathogen (OIE, 2009b). For ISAV detection, the targeted organ depends on the test method, the intended purpose of the test, and the stage of infection in the targeted animals. Organs should be

chosen for optimal detection methods based on information about the pathogenesis and route of host exposure.

Route of infection- The main route of transmission for ISAV appears to be horizontal (Melville & Griffiths, 1999). Past failure to experimentally infect fish *per os* (Totland et al., 1996) suggested that infection by coprophagy or cannibalism is less likely, although not ruled out completely (Mikalsen et al., 2001). Gills were suggested to be the most likely natural route of entry (Totland et al., 1996). Sea lice parasites and their associated skin lesions may play a role (Nylung et al., 1994). With a strong tropism for leucocytes and endothelial cells (Hovland et al., 1994), the virus starts replicating locally in the gills vascular endothelium, and by infecting deeper cell layers, penetrates lamellae and filament capillaries. According to the surveillance program in the Faroe Islands (Christiansen & Ostergaard, 2008), **gills** are the organ of choice to screen salmon for low-virulence genotypes (e.g. HPR-0), suggesting that this genotype might have a specific tropism for gills and might not spread further within the organism.

Viraemia and systemic spread- Once in the circulatory system, the virus has been shown to replicate in erythrocytes, leucocytes and endothelial cells (Hovland et al., 1994; Moneke et al., 2005; Workenhe et al., 2007) and spreads systemically. ISAV is described as a vascular disease where mortality and morbidity are due to severe internal haemorrhaging blood loss and subsequent hypovolemic shock. It takes an average of 7 to 13 days for the virus to induce repercussions on the vascular system, depending on the infective dose and the virulence of the agent. **Blood** has been identified as a potentially

non-lethal sample as supported by studies comparing this sample type with other tissues (Giray et al., 2005). Following the circulatory system of the salmon, ISAV is believed to spread rapidly into well perfused organs, namely heart, kidney, spleen, intestines, and liver. Due to the extensive network of endothelium comprising the **heart** stroma, this organ is currently considered the most appropriate target for early detection using sensitive methods such as RT-PCR (Mikalsen et al., 2001; Snow et al., 2003).

In the **liver**, infection of the vascular endothelium affects vascular integrity resulting in reduced blood distribution, generating severe congestion. With limited blood supply and oxygenation, the degeneration of hepatocytes is triggered resulting in multifocal to confluent regions of necrosis (Spielberg et al., 1995). Subsequently, increased portal venous pressure may cause the impediment of the sinusoidal and returning blood flow to the heart, often leading to the accumulation of ascites (Evenson et al.1, 1991). Ascites, however, may also be secondary to hypoproteinaemia caused by extensive liver damage, and severe nephropathy inducing decreased production of plasma protein (Simko et al., 2001). Severe nephropathy may similarly be associated with hypoproteinaemia (Simko et al., 2001).

In the **kidney**, virus affects mainly the posterior (excretory) component rather than the anterior (haematopoietic) segment of the organ (Simko et al., 2001). Interstitial vascularisation is affected, with the appearance of multifocal, sometimes coalescing to diffuse, haemorrhages (Byrne et al., 1998). As with the liver, this internal haemorrhaging will lead to both blood flow and oxygenation problems, as reflected by necrosis of tubular epithelium in some cases. Although possessing a similar blood supply, the anterior kidney seems to be less affected. Tubular lesions may be associated, caused by

peritubular haemorrhages. Necrotic foci have been observed in the hematopoietic tissues and are suspected to involve cells targeted by the virus, such as macrophages (Byrne et al., 1998). While prominent pathological changes were observed in the liver in Europe, kidney lesions were the primary lesions in early North American cases which erroneously led to the differentiation of the two pathologies: ISA in Norway, and haemorrhagic kidney syndrome (HKS) in Canada (Mjaaland et al., 2002). In practice, the kidney has been the sampling organ of choice.

Fixed macrophages are actively involved in the phagocytic activity in the renal portal endothelium and in other reticuloendothelial tissues, such as the **spleen** and kidney. Splenic lesions usually include diffuse congestion with degrees of erythrophagocytosis (Simko et al., 2001). Snow et al. (2003) experienced difficulty detecting the pathogen in spleen samples, suggesting that the splenic endothelium is not the primary location of virus replication. Depending on the individual fish, pathologic findings may also be observed in the intestines, pyloric caeca and stomach mucosa. Pathologic changes to these tissues can range from minor congestion and haemorrhaging within the lamina propria to severe mucosal necrosis and sloughing with marked intralumenal hemorrhage (Simko et al., 2001, Godoy et al. 2008).

Excretion phase- After 7 days post-infection (dpi), the fish begins shedding virus into the surrounding environment (Gregory et al., 2009) through a variety of external products, such as mucus, faeces and urine (Totland et al., 1996). Two to three days before mortality (approximately 14 dpi), the shedding is maximal, associated with increased probability of transmission to surrounding fish (Gregory et al., 2009). At this late stage,

Griffiths & Melville (2000) identified **gill mucus** as more likely to be positive for virus, compared to serum, using RT-PCR as a non-lethal detection method.

Host response to infection- A progressive decrease of the viral load was observed after 15-20 dpi, while a minimum of 25 dpi is required before the first evidence of a host immune response is apparent (Milkalsen et al., 2001). Resistance to re-infection was shown in fish previously infected or after passive immunization (using serum from recovering salmon), suggesting the presence of an adaptative humoral immunity against ISAV (Falk & Dannevig, 1995). Evidence of two types of antibody responses has been reported (Kibenge et al., 2002). Specific antibody responses were observed in recent acute infections, while cross-reacting antibody responses were suggested in chronic infections or resistance to ISAV. Subsequent production of **immunoglobulin** in response to ISAV infection represents an indirect analytical target for detection of infection in serum samples (Kibenge et al., 2002). Degrees of protection were also reported in salmon vaccinated with inactivated ISAV (Jones et al., 1999, Brown et al., 2000, Salonius et al., 2003) and may lead to false positive results.

Although, proper tissue selection and sensitive detection of ISAV is challenging in the early stages of infection, the exponential replication of the virus and the subsequent development of systemic infection generates several options for specimen collection. In case of suspicion, subsequent visits to the production site should increase the chance of detecting the virus if present. Persistently low levels of infection would be surprising for virulent genotypes of ISAV. Asymptomatic (i.e. HPR0) or healthy carriers of the virus

may, however, exist and represent greater challenges with the test methods currently available.

1.3.3 Developed ISAV diagnostic methods

Although Norway experienced ISA episodes as early as 1984 (Thorud & Djupvik, 1988), the viral agent was first visualized by electronic microscopy (Hovland et al., 1994) and grown in cell culture (Dannevig & Falk, 1994) in 1994. Prior to this time, diagnosis of ISA was conducted using gross pathology, haematology and histopathology (Evensen et al., 1991). Virus isolation (VI) was thereafter developed and standardized by Dannevig et al. (1995a) in well established cell line salmon head kidney 1 (SHK-1) and later in other salmonid cell lines, including: Chinook salmon embryo (CHSE-214; Bouchard et al., 1999), Atlantic salmon kidney (ASK; Devold et al., 2000), salmon head kidney leukocytes (TO; Wergeland & Jakobsen, 2001), Atlantic salmon (AS; Sanchez et al., 1993), and epitheliocytes of carp (EPC, Godoy et al., 2008). Identification of ISAV propogated by cell culture is achieved using adjunct assays. Initially confirmed by electron microscopy, the presence the viral agents was followed later by verification using mouse monoclonal antibodies against the HE surface protein in an indirect immunofluorescent antibody test (IFAT) (Falk & Dannevig, 1995).

Currently, monoclonal antibodies against ISAV are also used to detect and localize viral particles in tissue sections using immunohistochemistry (Falk et al., 1998), or in commercial field test kits (Aquatic Diagnostics Ltd., Stirling, Scotland). VI remained the main method of detection until 1997 when a nucleic acid amplification test

(NAAT) method was developed using reverse transcriptase-polymerase chain reaction (RT-PCR) targeting the RNA segment 8 of the viral genome (Mjaaland et al., 1997). Subsequently, alternative versions of RT-PCR were developed, including: one-tube (Melville & Griffiths, 1999), nested (Løvdal & Enger, 2002) and real-time (Munir & Kibenge, 2004). RT-PCR has also been used as an adjunct method to identify ISAV in CPE positive cell cultures. The sensitivity and resolution of localizing viral particles in infected tissues was later improved using riboprobes for *in situ* hybridization (ISH) (Gregory, 2002; Moneke et al., 2003). Finally, indirect diagnostic methods have recently been developed to detect antibodies against ISAV in infected and/or vaccinated salmon sera with direct or indirect competitive enzyme-linked immunosorbent assay (ELISA) (Kibenge et al., 2002). A summary of ISAV detection methods is presented in Table 1.5.

1.3.4 Past evaluations of ISAV diagnostic tests

Justified by the importance to accurately detect ISAV for control and surveillance programs, the "field" efficacy of common ISAV detection assays has been evaluated in several studies. McClure et al. (2005) evaluated the performances of histopathology, VI, IFAT and RT-PCR using a standard of reference, while Nérette et al. (2005a, 2008b) did not include histology and Gustafson et al.(2008) focused on IFAT and RT-PCR using advanced analytical methods without reference standard information (i.e. a latent class model).

Table 1.5. Review table of developed diagnostic procedures for Infectious Salmon Anaemia Virus.

	Targeted organ or sample	Analytical target	Diagnostic Target	Development & optimization references	Evaluation references
Field-based diagnostic methods					
Behavioural changes*	na	Mortality	Dis.	Gustafson et al.[2005]	
Clinical signs*	Eye, skin, fins, vent, gills	Hemorrhagic lesions	Dis.	NORWAY Thorud & Djupvik[1988], Thorud[1991], Evensen et al.[1991] CANADA & USA Mullins et al.[1998], Byrne at al.[1998], Bouchard et al.[2001] UK, FAEROE ISLAND Rodger at al.[1998],Anonymous[2000] CHILE Godoy et al.[2008]	
Clinical-based diagnostic methods					
Gross pathology*	Peritoneal walls, liver, spleen, intestinal wall, kidney, skeletal muscle	Hemorrhagic lesions, congestion, necrosis	Dis.	Hovland et al.[1994], Falk & Dannevig[1995], Spielberg et al.[1995], Byrne et al.[1998]	Opitz et al.[2000]
Haematology*	Blood	Hematocrite, Corticol, Lactate, Glutathione, Total plasma protein. Hyperglycaemia	Dis.	Hjeltnes et al.[1992], Olsen et al.[1992], Speilberg et al.[1995], Simko et al.[2001]	Opitz et al.[2000]
Smears*	Blood		Dis.		
Microscopic pathology* (histology)	Gills, liver, spleen, kidney, heart		Dis.	Evensen et al.[1991], Speilberg et al.[1995], Byrne et al.[1998], Mullins et al.[1998], Simko et al., [2000]	Opitz et al.[2000], Snow et al., [2003], McClure et al.[2005]
Electronic microscopy*			Inf. or Dis.	Hovland et al.[1994], Nylung et al.[1995], Speilberg et al.[1995], Koren & Nylund[1997], Workenhe et al.[2007]	Nylung et al.[1995]
Analytical-based diagnostic methods					
Virus isolation*	Gills, liver, spleen, kidney, heart	Active viral particle[DIR]	Inf. or Dis.	Dannevig et al.[1993], Dannevig & Falk[1994]; SHK-1 Dannevig et al.[1995a,b], Dannevig et al., [1997]; CHSE-214 Bouchard et al.[1999], Lovely et al.[1999], Kibenge et al.[2000a]; ASK Devold et al.[2000], Rolland et al.[2003]; TO Wergeland & Jakobsen[2001]; AS Sanchez et al.[1993]; Sommer & Mennen[1996 & 1997]; Rtgill-W1 Bols et al.[1994], Falk et al.[1997]; EPC Godoy et al.[2008]	McAllister[1997], Falk et al.[1998], Kibenge et al.[2000a,b], Opitz et al.[2000], Grant & Smail[2003], Moneke et al.[2003], Rolland et al.[2003], Snow et al.[2003], Munir & Kibenge[2004], McClure et al.[2005], Rolland et al.[2005], Nérette et al.[2005a,b,2008b]
Indirect fluorescence antibody test*	Kidney smear, adjunct to VI (confirmatory)	Hemagglutinin Esterase[DIR]	Inf. or Dis.	Falk & Dannevig[1995], Mjaaland et al. [1997], Falk et al.[1998], Blake et al.[1999], Bouchard et al.[1999], Rimstad et al.[1999], McBeath et al.[2006]	Opitz et al.[2000], Snow et al.[2003], McClure et al.[2005] Nérette et al.[2005b, 2008b], Gustafson et al.[2008]
Lateral flow immnoassay	Fresh kidney tissue	Nucleoprotein[DIR]		Aquatic Diagnostics Ltd. (http://www.aquaticdiagnostics.com/)	
Immunohistochemistry*	Histology embedded tissue section	Hemagglutinin Esterase[DIR]	Inf. or Dis.	Falk & Dannevig[1995], Falk et al.[1998], Wilson et al.[2002], OIE[2009b]	
In situ hybridization	Histology embedded tissue section	Messenger RNA[DIR]	Inf. or Dis.	Gregory, [2002]; Moneke et al., [2003], Moneke et al.[2005], MacWilliams et al.[2007]	
Reverse transcriptase-polymerase chain reaction(RT-PCR)*	Gills, liver, spleen, kidney, heart, adjunct to VI (confirmatory)	Genomic RNA[DIR]	Inf. or Dis.	Mjaaland et al.[1997], Blake et al.[1999], Bouchard et al.[1999], Melville & Griffiths[1999], Rimstad et al.[1999], Devold et al.[2000], Kibenge et al.[2000a,b], McBeath et al.[2000], Mikalsen et al.[2001], Løvdal & Enger[2002], Mjaaland et al.[2002]	Opitz et al.[2000], Snow et al.[2003], McClure et al.[2005], Nérette et al.[2005a,b, 2008b], Gustafson et al.[2008]
Quantitative RT-PCR*	Gills, liver, spleen, kidney, heart	Genomic RNA[DIR]	Inf. or Dis.	Munir & Kibenge[2004], Plarre et al.[2005], Snow et al.[2006], Starkey et al.[2006], Snow et al.[2009]	Workenhe et al.[2008]
Loop mediated isothermal amplification	Gills, liver, spleen, kidney, heart	Genomic RNA[DIR]	Inf. or Dis.	McCarthy et al.[2006]	
Isothermal rolling circle amplification	Gills, liver, spleen, kidney, heart	Genomic RNA[DIR]	Inf. or Dis.	McCarthy et al.[2007]	
Haemadsorption	Adjunct to VI	Hemagglutinin Esterase[DIR]	Inf. or Dis.	Smail et al.[2000]	
Virus neutralization	Blood serum	Antibody anti-ISAV[IND]	Inf. or Dis.	Joseph et al.[2003], Kibenge et al.[2004]	
Enzyme-linked immunosorbent assay*	Blood serum	Antibody anti-ISAV[IND]	Inf. or Dis.	Kibenge et al., [2002], Falk [Pers. Com.]	Cipriano[2009]

*listed in OIE Manual (2009b)

[DIR] direct
[IND] indirect

Inf.: infection
Dis.: Disease

SHK-1: Salmon Head Kidney
CHSE-214: Chinook Salmon Embryo
ASK: Atlantic Salmon Kidney
EPC: Epithelioma papulosum cyprini
AS: Atlantic salmon
TO: Salmon head kidney leukocytes
Rtgill-W 1: Rainbow Trout gill

na: non applicable

1.4 Methodology to evaluate precision of dichotomous tests

According to international standards, the precision of a test method is defined as the agreement between independent test results obtained under prescribed conditions (ISO 5725-1, 1994). Precision does not necessarily account for the true value but rather reflects the distribution spread of random error of the test. Obviously, ISO definitions are more adapted for methods that measure continuous outcomes. For these methods, precision is usually expressed by estimates of the standard deviation (variability of repeated measures) or maximum expected difference between two measurements under specific conditions with 95% confidence.

Dichotomous tests yield binary outcomes (positive or negative) and the distinction between random and systematic error is not as straightforward. With a continuous test, a method can be precise (consistent measurements) and not true (biased). However, when results from two binary test results disagree, one must be true and the other false. Therefore, when a dichotomous test is not true, it is considered to be imprecise. For binary test results, (dis)agreement combines measures of (in)accuracy (systematic error), (im)precision (random error) (Van der Bruel et al., 2007) and dependence among sub-samples from a same specimen. A dichotomous test that does not yield consistent results experiences some degree of variation in its DSe and DSp. The degree of variation is closely associated with the conditions under which the test was run. Within this discussion, the terms “test runs” or “runs” are used for a set of results obtained using the same test method and set of samples under specified testing conditions defined in the study design. Here, we will refer to agreement as the comparison of results from the same

method under differing conditions. Estimation of agreement between different tests is beyond the scope of this section, although it follows similar principles.

1.4.1 Study design and associated bias

1.4.1.1 Design

The major consideration in the design of a study to evaluate result agreement is to clearly describe the test protocol. For instance, Nérette et al. (2005b) evaluated the consistency of an ISAV RT-PCR assay across 3 laboratories. Each laboratory used the same general technique but with different protocols (i.e. different amplification primers) which seriously impacted the comparability of laboratories and instead resulted in a comparison of separate methods. Therefore, the protocol of the studied test should be as precise and standardized as possible to only compare applications of the same protocol. The second main design consideration is to define the conditions under which the test is run and compared. Indeed, agreement is evaluated under two sets of conditions: (i) runs of test were repeated under identical conditions within a laboratory (repeatability); (ii) and runs of test were repeated under similar conditions between laboratories (reproducibility). Sources of variation due to the laboratory can be chronologically separated in 3 phases; briefly, the sampling procedure (e.g. targeted organ, pooling); the testing process (e.g. operator); and the test result interpretation (e.g. decision threshold) (OIE, 2009b). One important OIE objective is to, ultimately, ensure that a prescribed test can be transferred to a wide range of users with minimal change in performance (e.g. test

kits). In addition to repeatability and reproducibility, OIE guidelines require several other measures of agreement under specified conditions and design summarized below:

1- *Test robustness.* In the optimization phase of test development, the robustness of the method is estimated by the consistency of test results across a wide range of testing conditions within the same laboratory. This evaluation should be customized to fit the technical specificities of the method and focus on factors likely to influence the performance of the assay in question.

2- *Analytical repeatability.* The evaluation of analytical repeatability requires estimates of within-run agreement using at least two replicate blinded samples, and repeated at least 20 times (between-run agreement) on at least 5 separate days (between-day agreement) and with at least 2 different operators (between-operator agreement). The allocation of the 20 runs among days and operators is not specified. However, a logical allocation would be that each operator runs two sets of tests per day for five days.

3- *Analytical reproducibility.* Following of the analytical evaluation, a preliminary evaluation of analytical reproducibility with a small set of highly characterized samples is recommended. However, this phase primarily serves to obtain the provisional assay recognition at the end of the analytical evaluation (see section 1.2.1.2.1).

4- *Field reproducibility and repeatability.* Stage 3 of the evaluation is specifically dedicated to the evaluation of field reproducibility and extension to the field repeatability (also referred to as “field” evaluation of test precision). Using blinded aliquots, field samples are tested in duplicates at each of the participating laboratories (at least 3). Also referred to as test ruggedness, the influence laboratory factors (also called “user” factors,

Crowther et al., 2006) on test result agreement can be estimated by comparing within-laboratory variation with the variation among-laboratories.

5- *Internal and external proficiency testing.* To maintain assay validity, the proficiency of the laboratory (test) must be monitored and controlled through the implementation of internal and external quality control programs. Internally, repeatability of control samples is continuously monitored to detect temporal trends. Externally, biannual proficiency testing (e.g. ring test) is conducted to monitor the reproducibility across laboratories.

6- *Comparability Assessments*. When changes occur in either the structure of the target (e.g. new pathogen strain), assay reagents, or in the technology used for detection, it is necessary to demonstrate comparable performances instead of re-validation. Comparison of the previous and modified methods in parallel over at least 20 runs is recommended providing an evaluation of agreement between the two protocols.

In summary, the design to study dichotomous test agreement requires numerous aliquoted specimens tested repeatedly by blinded operator(s) under specified conditions. Depending on the factor of interest, replicated samples can be randomly allocated within run to reduce systematic bias. For instance, within a test plate, the samples can be randomly allocated to avoid effect such as plate margins.

1.4.1.2 Sampling considerations

The sampling strategy to evaluate test precision depends on numerous factors associated with the target population, disease of concern and test technique. Although it

is difficult to provide a generic sampling protocol applicable in all situations, OIE imposes some requirements regarding the sample size and the origin of the samples.

1.4.1.2.1 Number of samples

At least 6 estimation exercises involve evaluation of agreement to validate a dichotomous test.

1- *Test robustness.* OIE does not impose a minimum sample size to evaluate robustness. However, the power of this evaluation depends on the strength of association of specific factors with the technique consistency. Therefore, the required sample size depends on the method and the factor of concern.

2- *Analytical repeatability.* OIE requires a minimum of 3 samples representing the analyte activity within the linear range of the assay (OIE, 2009b). Duplicate samples must be tested in a minimum of 20 runs, assuming each of 3 samples is aliquoted at least 40 times. There is no specific requirement that any of the 3 samples be analyte-free (i.e. negative controls). The choice of the three samples' concentrations is usually made subsequent to the estimation of the analytical sensitivity (limit of detection). Non-infected samples can be alternated with infected samples to enable agreement for non-infected samples to be estimated, and to investigate potential cross-contamination using "sentinel" samples.

3- *Analytical reproducibility.* OIE guidelines lack detail for requirement for this characteristic except to mention the need for a small panel of samples. Since it is a

preliminary evaluation, the same samples employed for the previous evaluation of analytical repeatability are often used.

4- *Field reproducibility and repeatability.* OIE requires a minimum of 20 field-derived samples be tested twice in at least 3 laboratories (OIE, 2009b). Each sample must be aliquoted at least 6 times and there is no specific requirement that any samples be analyte-free (i.e. proportion of infected samples). It would be useful to include half non-infected and half infected samples. The infected samples must represent the natural range of infection stage in the population. Within each run, samples can be randomly allocated or alternated between infected and non- infected specimens.

5- *Internal and external proficiency testing.* OIE does not provide guidelines to conduct quality control programs. Internal programs are normally conducted by close monitoring of results from control samples (one non-infected and one, or two, infected controls) included when running routine samples over time. Complementarily, external quality programs, or ring tests, are conducted periodically (e.g. twice a year) using the same template as the field evaluation for reproducibility and repeatability estimation.

6- *Comparability Assessments.* No specific sample size is required in each of the 20 runs where both methods are compared. The evaluation of comparability must be done using samples submitted in routine operations and tested by both methods.

In general, sample size calculations are directed by the expected magnitude and the degree of precision of the estimate. If agreement is expressed as a simple proportion of paired test results that agree, the sample size calculation for an expected proportion P_0 (between 0 and 1), estimation error range L (half of the expected confidence interval) and

type I error α (complement to confidence level), using a normal approximation, is (Dohoo et al., 2009):

$$n = Z^2_{(1-\alpha/2)} P_0 (1-P_0) / L^2 \quad (30)$$

where Z is the z distribution value associated with a confidence level associated with the coverage probability 1- α/2 (usually 97.5% for α = 5%).

1.4.1.2.2 Nature and origin of samples

The principal objective when selecting samples for test evaluation is that the samples are as close as possible to the ones that will be submitted during routine use. Both control and reference materials are used during the analytical evaluation of agreement, comprising robustness, ASe, analytical repeatability & reproducibility, and internal proficiency testing. The nature of reference samples (standards) varies from natural isolates to synthesized oligonucleotides (e.g. recombinant plasmid) (Wong & Medrano, 2005). The complexity of the sample matrix has a strong influence on the operating characteristic of the assay (Hiney & Smith, 1998). OIE recommends the use of field samples derived from the target population (i.e. the population in which the test will be used). Samples of various concentrations (e.g. low level of virus) may not be readily available from the field requiring the utilization of serial dilutions that may impact the structure and complexity of the sample matrix. The addition of a complementary fraction of a neutral solution (e.g. molecular graded water) to a concentration of analyte will

artificially dilute the matrix without adjusting for host and/or environmental material. Thus, OIE requires diluting highly infected samples with analyte-free matrix from hosts in the target population (OIE, 2009b). Spiked samples are not recommended since they do not represent biological reality, albeit they may be useful when infectious material is not available.

Samples included in the evaluation should be representative of the spectrum of infection (and biological factors associated with the expression of the disease) encountered in the current or future target population. For accessible populations, there are a wide variety of sampling strategies used to collect real specimens, each potentially associated with different degrees of selection bias. These strategies are the same as for samples used in test accuracy evaluation and therefore reviewed further below (see section 1.5.1.2.2*)*. Although similar samples may be used for the two types of studies for field reproducibility and repeatability estimation, a pool of samples with a medium range prevalence of infection (e.g. 50%) is recommended to limit problems during Kappa estimation (Chapter III for further explanations).

Once collected, the main concerns about reference and control materials are the *homogeneity* among sub-aliquots and the *stability* of the specimen (i.e. storage) (Van der Veen & Pauwels, 2000). Indeed, to fit repeatability and reproducibility conditions, it is critical for replicated samples to be identical. For instance, Nérette et al. (2005b) used separate tissue samples dissected from the same salmon kidney sent to 3 different laboratories for testing. It was, thereafter, argued that observed discrepancy in test results among laboratories may have been due to heterogeneous distribution of ISAV particles among sub-samples and not due to laboratory performances. Although agreement may

have been underestimated, the random allocation of the sub-samples ensured that one laboratory was not systematically favoured over another. The development of sample processing techniques, such as tissue homogenization, is needed to ensure qualitative reference materials for test precision and proficiency studies. Albeit essential for evaluation of test precision, methods to assess the homogeneity and stability of control samples is beyond the scope of this discussion and are covered in detail elsewhere (Van der Veen et al., 2000; 2001a&b). When replacing reference samples (or reagents), it is advised to overlap the use of both materials (i.e. original and replacing materials), for a number of runs to ensure comparability (ISO/IEC 17025, 2005).

1.4.1.3 Bias

For repeatability studies, duplicate samples are used. It is important that the operator is blinded to avoid any information bias. For test accuracy studies, Ransohoff & Feinstein (1978) referred to this bias as the review bias. Duplicate samples should be coded to remove source identification and randomly allocated to the testing order. Mislabelling or contamination should be suspected when a large majority of the laboratories detected target from supposed non-infected samples (Crowther et al., 2006).

1.4.2 Analysis methodology[2]

This section reviews the different ways to express and compute agreement including the statistical methods available to compare agreements.

[2] This section was adapted from the webpage "Statistical methods for rater agreement" edited by John Uebersax in 2007: www.john-uebersax.com/stat/agree.html.

1.4.2.1 Agreement parameters (Proportion of agreement / Kappa)

There are numerous ways to report test result variation. For tests with binary outcomes, the concept of test result variation is traditionally expressed as an estimation of agreement between test runs. Cleophas et al. (2008) suggested that precision for binary tests can also be expressed using predictive intervals (random variation) of DSe, DSp or Ef. However, OIE guidelines explicitly require measures of agreement (OIE, 2009b). Conventionally, agreement is expressed as proportion of agreement (Pa) (i.e. proportion of test results that agree), or as Cohen's Kappa values (κ) (Dohoo et al., 2009), but other more refined parameters exist. For instance, the **tetrachoric correlation** is another described way of measuring agreement. Developed by Pearson (1900), this statistic makes 6 assumptions that cannot be tested with only two test runs (Uebarsax, 2007). Therefore, this measure is rarely used.

To illustrate agreement parameters, we will use the example of two test runs (1 & 2) from paired samples reported in a classical 2X2 contingency table:

		Test run 2		
		Positive	*Negative*	
Test run 1	*Positive*	*a*	*b*	*a + b*
	Negative	*c*	*d*	*c + d*
		a + c	*b + d*	*n =a + b + c + d*

Proportion of agreement (Pa) – Also called *observed agreement*, Pa represents the overall proportion of paired test results that agree and is computed as:

$$Pa = (a + d) / (a + b + c + d) = (a + d) / n \quad (31)$$

The standard error of Pa (SE_{Pa}) is calculated as any other proportion ($SE_{Pa} = \sqrt{[Pa*(1-Pa)/ n]}$) and confidence intervals can be computed using a normal approximation, an exact binomial distribution or bootstrap or Jackknife approaches. Pa has the advantage of being simple to compute and interpret. However, it does not distinguish between agreement for positive versus negative test results and does not correct for the proportion of results that agree only by chance (random classification).

Positive and negative agreement (Pa+ & Pa-, respectively) – Also called *proportions of specific agreement*, Pa^+ and Pa^- are computed as follow:

$$Pa+ = 2a / (2a + b + c) \quad (32)$$

$$Pa- = 2d / (2d + b + c) \quad (33)$$

In Eq. (32), *a* replaces *d* from Eq. (31) to only count for agreement on a positive result. Eq. (33) is the reciprocal. Asymptotic standard errors may be calculated using Mackinnon's formulae, bootstrapping or Jackknife estimation (Uebersax, 2007). Confidence intervals can be obtained by normal approximation, bootstrap or Jackknife approaches (not binomial). Assuming infected samples increase Pa+ and non-infected samples decrease Pa+, the difference between Pa+ and Pa- reflects the prevalence of infected samples in the tested pool, and therefore this difference gives an idea of the inflation of agreement due to chance. For instance, if the infection prevalence is very low in the sample pool and inflation of agreement due to chance is present, Pa- will be very

large (i.e. more inflated) compared to Pa+. Assuming independence, agreement due to chance is exacerbated in extreme prevalences and tends to overwhelm the intrinsic agreement due to the test. To obtain an unbiased estimation of agreement, Cohen (1960) developed a Kappa index that accounts for agreement due to random classification.

Cohen's Kappa coefficient (κ) – κ is defined as the ratio of the *actual agreement* beyond chance over the *potential agreement* beyond chance. The actual agreement (the numerator) is the difference between the observed agreement (Pa) and the *expected agreement* (due to chance). The potential agreement is the difference between the maximal agreement (i.e. 1) and the expected agreement. The components of the κ computation are illustrated in Appendix 3. The expected agreement corresponds to the agreement by chance alone and is computed using the marginal probabilities from the 2X2 table assuming conditional independence between tests:

Expected agreement =[(a+b)(a+c) / (a+b+c+d)] + [(c+d)*(b+d) / (a+b+c+d)] (34)*

This computation of agreement by chance alone assumes that the two test runs are statistically independent. However, this assumption is rarely true and the computed expected agreement may not be valid. Therefore, κ is not really considered as a chance-corrected measure of agreement (Guggenmoos-Holzmann, 1996). Nonetheless, κ measures dependence between test runs (or the deviation from independence) (Gardner et al., 2000). Significant evidence of κ being > 0 confirms that observed agreement exceeds agreement by chance and that the test runs are conditionally dependent. Conversely, if

there is no significant evidence of κ being different from 0, it is confirmed that the observed agreement is not different from the agreement by chance and that the runs are conditionally independent. In the κ calculation, the computed agreement by chance corresponds to the agreement of conditionally independent test runs and is reciprocal. Test runs are considered completely dependent when κ = 1. The computation of κ can be simplified as (Dohoo et al., 2009):

$$\kappa = 2\,(a\,d - b\,c) \,/\, [(a+b)*(c+d) + (b+d)*(a+c)] \qquad (35)$$

Procedures to compute the standard error, confidence interval and test of significance are reviewed elsewhere (Fleiss et al., 2003).

Several variations of Cohen's κ exist (e.g. Fleiss's Kappa) and are described elsewhere (Uebersax, 2007). Most the statistical packages compute the different variants of κ.

Utilization of κ is associated with numerous controversies and should not be used as the only estimation of agreement (Maclure & Willet, 1987). Factors influencing κ include DSe and DSp of each test run, sensitivity and specificity covariance between test runs, and the prevalence of infection (Gardner et al., 2000). The strong dependence on prevalence is one of two paradoxes of κ (Feinstein & Cicchetti, 1990). Similar to Pa, κ depends on prevalence but is not a prevalence-weighted average of a specific κ in D+ and D- individuals (i.e. not a linear function of prevalence). With extreme prevalences, κ tends to drop towards 0 (Byrt et al., 1993). It is therefore not recommended to evaluate agreement for extreme prevalences (< 20% and > 80%, Dohoo et al., 2009), but instead to

use a set of samples with a medium range prevalence. The second paradox describes the fact that κ is overestimated when the proportion of positive test results (marginal proportions) differs between test runs (Feinstein & Cicchetti, 1990). Before κ estimation, a test of *marginal homogeneity* (e.g. McNemar's $\chi 2$, see below) is required to investigate the presence of a bias. Estimated κ values between test runs with significantly different proportions of positive test results are biased and of little interest.

1.4.2.2 Statistics: McNemar's / Symmetry / Marginal homogeneity test

Computation of κ is only worthwhile if there is no obvious evidence of disagreement between test runs (i.e. proportion of positives differing strongly). Several procedures exist that are equivalent or synonymous in the 2x2 setting (i.e. 2 runs/2 outcomes). Usually referred to as McNemar's test, the hypothesis of the two runs yielding equal proportions of positives (H_0) can be approached as the test of marginal homogeneity or the test of symmetry. In the follow 2x2 contingency table:

		Test run 2		
		Positive	*Negative*	
Test run 1	*Positive*	p_{++}	$\boldsymbol{p_{-+}}$	$p_{.+}$
	Negative	$\boldsymbol{p_{+-}}$	p_{--}	$p_{.-}$
		$p_{+.}$	$p_{-.}$	*1*

the marginal probabilities (in grey) are the proportions of respective results when each test run is considered separately. The symmetrical cells (in bold) correspond to the proportions of disagreement symmetrically distributed across the diagonal of agreement cells (not bold). For 2x2 settings, the symmetry condition implies marginal homogeneity

and vice versa (Pendergast et al., 2005). McNemar's test can therefore be conducted both ways, either by computing the McNemar's χ^2 statistic (marginal homogeneity) or the exact binomial statistic (for symmetry) (Dohoo et al., 2009). The McNemar's χ^2 is computed as follows (Dohoo et al., 2009):

$$McNemar's\ \chi^2 = (b-c)^2 / (b+c) \qquad (36)$$

The P-value is obtained from the corresponding cumulative probability of χ^2 distribution with 1 degree of freedom. The exact binomial test is based on the null hypothesis that if the marginal probabilities were equal (H_0), the disagreement cells should be equal and the probability to be in one or the other should be also equal (i.e. 50%), Therefore, the null hypothesis can be tested by computing the probability of the observed number of disagreeing samples in one cell (k) out of the total disagreeing samples (n) assuming that the binomial probability of either cells was 50%:

$$Prob(X:k \mid p=.5) = C^n_k\ 0.5^k\ 0.5^{n-k} \qquad (37)$$

This probability corresponds to the probability to accept H_0 when it is not correct (*P*-value). The alternative hypothesis supports that a significant difference exists between the proportions of positives between test runs. If this is the case, then serious disagreement exists between runs, making the estimation of κ of little value (Dohoo et al., 2009).

Estimates of agreement can be obtained under different conditions (e.g. repeatability) and these agreements can be compared (e.g. repeatability in different

laboratories). Agreement comparisons involving the paired test runs (same samples tested under different conditions) are specifically addressed in this section. However, simpler procedures, such as a comparison of multinomial distributions, can be used when test run results are independent (not paired). The results from each agreement estimation yields 4 different combinations of test results (4 inner cells of the 2X2 table) summarized in a cubic 4X4 contingency table as follows:

	Result combination	*Condition 2*				
		+/+	+/-	-/+	-/-	
Condition 1	+/+	$p_{+/+,+/+}$	$\boldsymbol{p_{+/-,+/+}}$	$\boldsymbol{p_{-/+,+/+}}$	$\boldsymbol{p_{-/-,+/+}}$	$p_{.,+/+}$
	+/-	$\boldsymbol{p_{+/+,+/-}}$	$p_{+/-,+/-}$	$\boldsymbol{p_{-/+,+/-}}$	$\boldsymbol{p_{-/-,+/-}}$	$p_{.,+/-}$
	-/+	$\boldsymbol{p_{+/+,-/+}}$	$\boldsymbol{p_{+/-,-/+}}$	$p_{-/+,-/+}$	$\boldsymbol{p_{-/-,-/+}}$	$p_{.,-/+}$
	-/-	$\boldsymbol{p_{+/+,-/-}}$	$\boldsymbol{p_{+/-,+/+}}$	$\boldsymbol{p_{-/+,-/-}}$	$p_{-/-,-/-}$	$p_{.,-/-}$
		$p_{+/+,.}$	$p_{+/-,.}$	$p_{-/+,.}$	$p_{-/-,.}$	*1*

Although less informative, an alternative approach is to group the resulting combinations into two categories: agreeing results (+/+ & -/-) and disagreeing results (+/- & -/+). The resulting 2X2 table can then be analysed and tested with the McNemar's test. In this instance, only the magnitudes of agreement are compared regardless of the nature of the result. Alternatively, when specific agreement proportions are of interest, the resulting combinations can be rearranged in three categories (agree on a positive result, agree on a negative result and disagree). A test of symmetry can be conducted to compare symmetrical cells (bold). Asymptotic and exact computations of the symmetry χ^2 statistic exist and include all pairwise comparisons of symmetrical cells. However, marginal homogeneity does not prove symmetry in a cubic table larger than 2X2 (Pendergast et al. 2005). Tests for marginal homogeneity of several categories can also be implemented, with Stuart-Maxwell being the most commonly used (Stuart, 1955; Maxwell, 1970). More advanced procedures using quasi-symmetry model comparisons exist and are

described elsewhere (Agresti, 2002). However, it is not legitimate to compare two agreement estimates involving the same test run. For instance, when comparing reproducibility among 3 laboratories (3 pairwise comparisons among laboratories), the results of the comparison between 2 laboratories can be deduced from the 2 others pairwise comparisons. Therefore, the 3 pairwise estimates of reproducibility are not independent and cannot be tested.

In summary, the precision of a dichotomous assay reflects the consistency of the test classification. If a test is not sufficiently robust or rugged, the respective evaluation of repeatability and reproducibility are of little value. Nonetheless, a test can be repeatable and reproducible even when it is inaccurate. The next evaluation stage requires assessing the efficiency of the test to properly classify samples. Later on, continuous monitoring of agreement will ensure consistent performance over time (internal quality control) and demonstrate equivalency with other methods or laboratories (external quality control).

1.5 Methodology to evaluate trueness of dichotomous tests

Often mistaken with accuracy, the trueness of a test is defined as the closeness of agreement between the average value obtained from a large series of test results and an accepted reference value (ISO 5725-1, 1994). Trueness relies on the true value and reflects the test measurement bias or systematic error (difference between the expected test results and an accepted reference value). The combination of precision (complement

of random error) and trueness (complement of systematic error) results in the estimation of accuracy defined as the closeness of agreement between a test result and the accepted reference value (ISO 5725-1, 1994). Better adapted to tests with continuous outcomes, the definition of trueness could be refined for dichotomous tests as the proportion of values obtained from a large series that agree with the accepted reference status. As describe previously (see section 1.2.1.2.2), this proportion corresponds to the efficiency (Ef) of the test. However, Ef depends on prevalence and may vary substantially across tested populations (Alberg et al., 2004). Other parameters that express overall trueness have been considered (e.g. Youden index (J), diagnostic odds ratio (DOR)), and they can be computed from diagnostic sensitivity (DSe) and specificity (DSp) (infection/disease status specific parameters) (see section 1.2.3). This section will focus on the evaluation of DSe and DSp as indicators of the trueness of a diagnostic test. The validity of the estimation of DSe and DSp lies primarily on the study design and can be affected by many variables, particularly when applying such studies to field situations.

1.5.1 Study design and associated bias

The approach and the study design for diagnostic test evaluation are context dependent and cannot always fit in a rigid standard framework. However, the default design of test trueness studies requires that the *candidate* (or *index*) *test* be compared to at least one *reference* (or *relative*) *test*, both run on the same samples (*paired design*). The only cases of studies where a minimum of two methods is not required are **experimental infection challenges**. This type of study does not represent the natural range of infection

stages (*spectrum bias*; Ransohoff & Feinstein, 1978) expected in a true population (*selection bias*, Greiner & Gardner, 2000b) and therefore tends to over-estimate DSe and DSp (Greiner & Gardner, 2000b). As for agreement studies, OIE recommends to not use this approach (OIE, 2009b) for similar reasons. A range of designs exist associated with different evaluation objectives related to the intended utilization of the test (Fig. 1.5).

1.5.1.1 Design

The first consideration to address in diagnostic study design is the evaluation of efficacy and usefulness of a candidate test (Fig. 1.5). Partly addressed by the concept of "fitness for purpose", the evaluation of a diagnostic method may not require the exclusive estimation of absolute parameters (DSe/DSp) and may be based on the assessment of relative criteria associated with test usefulness and efficiency. For instance, the success of the implementation of the assay can be estimated by the clinical outcome (e.g. survival of a disease) or by the cost-effectiveness (e.g. control program cost) of the test (Van den Bruel et al., 2007). When the choice of available techniques is limited, the operator can choose by default the method that gives the least compromised consequences. Model estimations using information from comparative studies (e.g. clinical field trial) may identify the tests associated with the best outcome. In these settings, the patient is only tested with one assay and the comparison between two testing strategies involves two different sets of individuals (*unpaired design*). Except in this particular case, diagnostic evaluations usually use a *paired design* where the compared methods are tested on the same individuals. For instance, a new test proposed to replace an existing one can be

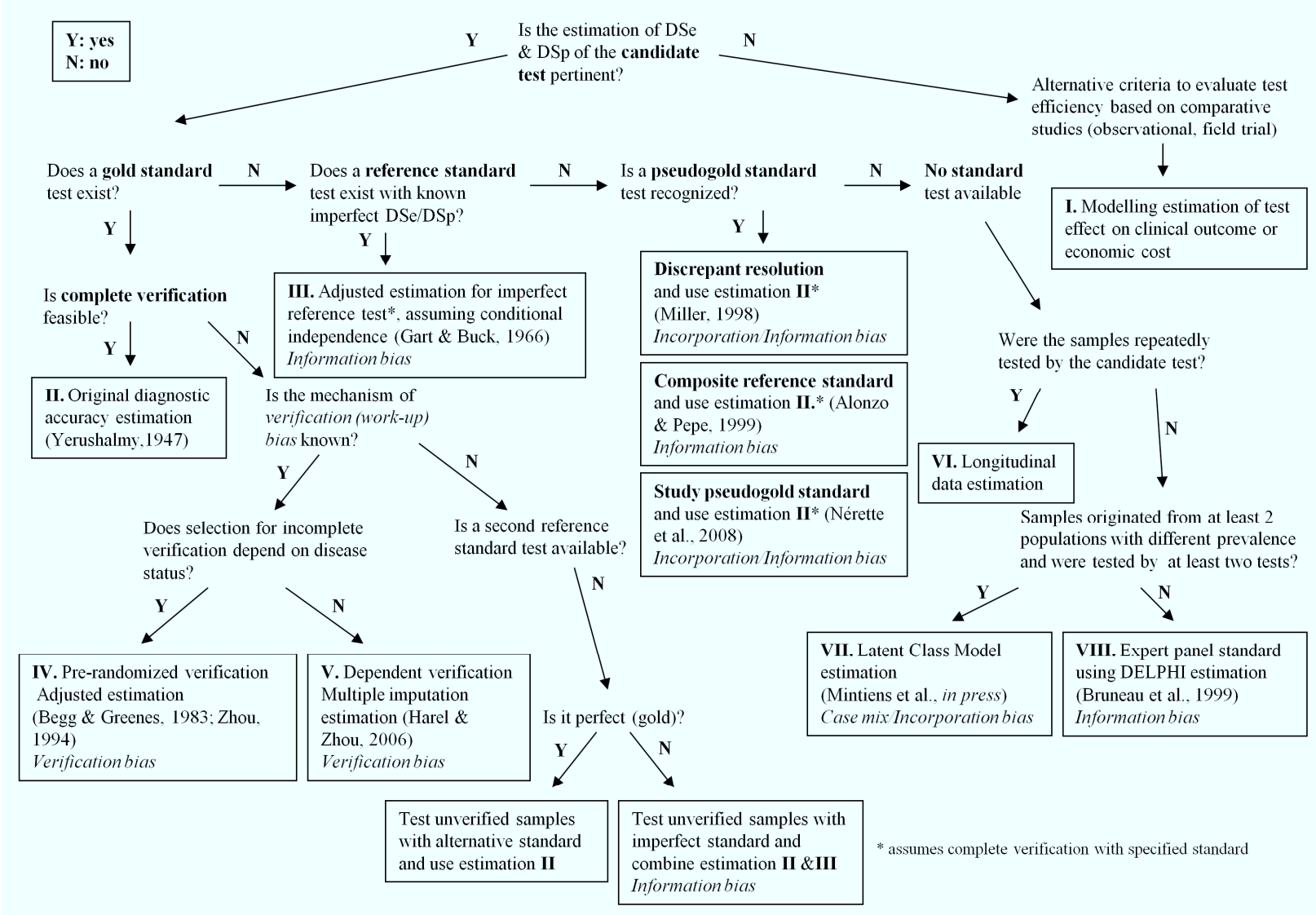


Fig 1.5. Decision tree to select estimation procedures for test trueness (adapted from Reitsma et al., 2009). Potential source of bias are presented in italics for each estimation procedure.

evaluated based on its relative classification performance (i.e. agreement). This agreement approach is mainly used to demonstrate *test comparability* where the candidate test is expected to be at most as accurate as the reference test. The relevance of a new test could therefore be justified on its reduced invasiveness, processing time, or cost.

Candidate tests may perform better than the reference test, and in such situation, the relative values of DSe and DSp can be estimated using a latent class model based on results of samples from populations with different prevalences (Bertrand et al., 2005). In this case, relative estimates of DSe and DSp are of interest for test comparisons. Overall, absolute estimates of DSe and DSp appear to be the most commonly used criteria to evaluate and interpret a test, and the international community and institutions adopted them as standards for test evaluation, validation and certification (OIE, 2008).

The second design consideration for trueness estimation is the type of standard information available to compare the candidate test (Fig. 1.5). In this discussion, *standard* refers to the reference test(s) or information used to identify the true infection/disease status of the sampled individuals used in the study.

The first standard to consider is the **gold standard** (GS). By definition, the GS provides the correct classification of the studied specimen. A GS test is deemed perfect (100% DSe and DSp). When a GS is available, the estimation of DSe and DSp of the candidate test is straightforward (see section 1.5.2.1) and follows the original definitions (Yerushalmy, 1947). However, the existence of perfect standard tests raised concerns among experts (Greiner & Gardner, 2000b) considering that such standard is not truly

gold but instead “silver, bronze or even tin” and might have led to biased estimations (McKenna & Dohoo, 2006).

When the reference test does not perfectly classify samples (**imperfect reference standard**, IRS), adjustment for *information bias* is possible (Greiner & Gardner, 2000b) based on known estimates of the IRS DSe and DSp, if available (Dohoo et al., 2009). This estimation procedure assumes that the two tests are conditionally independent and requires that all samples were tested by both tests (Enøe et al., 2000). The assumption of independence might not always be satisfied and, if rejected, requires more complex approaches (Hui & Zhou, 1998).

Not all samples are verified by supplementary tests in practice due to the invasive nature of the procedure or to the high associated cost. Therefore, only a fraction of specimens are tested twice, especially in clinical settings, which results in *verification* (or *work-up*) *bias* (Ransohoff & Feinstein, 1978), requiring corrective procedures. If the protocol of partial verification was conducted such that specimens were verified independently of their candidate test result (i.e. randomly selected), the estimation can be **directly adjusted** (Begg & Greenes, 1983; Zhou, 1994). If the proportions of verified samples are different in positive and negative result samples to the candidate test, the verification is dependent and requires complex **multiple imputation** procedures to adjust the estimation (Zhou, 1993; Zhou, 1998; Harel & Zhou, 2006). For instance, for a study population with a low prevalence of infection, the few positive results are usually all verified with the reference test while only a fraction of the negatives are randomly retested (Reitsma et al., 2009). Since, no formal method exists to test for verification

independence unless it was pre-specified, utilization of multiple imputation estimation is the default method of choice (Reitsma et al., 2009).

An alternative is to test thereafter unverified samples using a second reference standard test that is either less invasive or less expensive. In the very unlikely case of both reference tests being perfect, direct estimation of a trueness parameter can be conducted using estimation procedure used for a GS. However, if the second or both tests are imperfect, but the respective DSe/DSp are known, estimation can be conducted by combining the GS and/or imperfect test estimation procedures. Finally, if both reference tests are imperfect with unknown operating characteristics, a pseudo-gold standard approach is warranted, as described later. Overall, it is not recommended to use partial verification study designs.

When the operating characteristics of the reference test are unknown, it may be possible to use a **pseudo-gold standard** (PGS) (Fig. 1.5). The PGS classifies the specimen based on a combination of information usually obtained from at least two imperfect tests (Dohoo et al., 2009). All samples are thereafter considered correctly classified, and parameter computations are the same as for estimation with a GS. Three approaches have been described for defined PGS criteria. The first two use a preset framework using supplementary test results and the third relies on additional information.

The **discrepant** (or discordant) **resolution** is a two-stage process to classify samples (Miller, 1998a). All samples are first subjected to the candidate test and the reference test, assuming they are conditionally independent. Only the samples that yield discordant results (combinations of positive/negative test results) are tested in a second stage using a third test to resolve the discrepancy. If the first reference test is assumed

100% specific, only samples with positive candidate results and negative reference results are retested. Conversely, if the reference test is assumed 100% sensitive, only samples with negative candidate results and positive reference results are retested. In this second stage, it is possible to use several tests sequentially to increase the degree of confidence (Hadgu, 1996). Even with a perfect test used in the second stage, the DSe and DSp are always overestimated (Miller, 1998b). This is explained by the fact that, in the first stage, the candidate (imperfect) test is involved in the classification of the samples where the specimen status is deemed true when both tests agreed. However, some tested samples with agreeing results on both tests may in fact be false. Discrepant analysis combines *incorporation bias* (inclusion of the candidate test in the sample classification, Ransohoff & Feinstein, 1978) and *information bias* (misclassification of the samples for estimation). In its statistical guidance for diagnostic studies, the US Food and Drug Administration does not use discrepant resolution, justifying that "it does not solve a bias problem and it is a more complicated, incorrect solution" (FDA, 2007).

The second approach is referred to as **composite reference standard** (CRS) where the sample classification relies on a predefined combination of several test results that do not include the candidate test (Alonzo & Pepe, 1999). Usually at least two references tests are used and according to their expected characteristics and conditional (in)dependence, the interpretation of the combined results are done either in parallel or in series (see section 1.2.3.9). The combination of the references tests is therefore believed to increase the classification performances more than when used individually. Briefly, if the intent is to optimize specificity of the CRS, only samples testing positive with the first reference test are retested by the second test (or resolver test). Then, only samples that

yield two positive tests are deemed positive (i.e. series interpretation). If the intent is to optimize sensitivity of the CRS, only samples testing negative with the first reference test are retested by the second test. Thereafter, only samples that yield two negative tests are deemed negative (i.e. parallel interpretation). Sequential testing (i.e. only a fraction of samples tested by the resolver) is not a design requirement but is, in this instance, recommended for optimization of analysis cost. More than two tests might be used to define a CRS according to the objective of classification. However, improvement of the classification is ensured only if the tests are conditionally independent, and perfect classification may still not be achieved, implying information biased estimates. In the evaluation of three ISAV tests, Nérette et al. (2008b) used CRS to estimate DSe and DSp. The authors compared two of the tests and used the third one as a resolver. Interpreting the results of the supplementary tests in parallel, the estimation of DSe was heavily dependent on the resolver positive results, with estimates changing as much as 23% depending on which test was considered the resolver (Nérette et al., 2008b). The test with the best assumed DSp was selected as the reference test (i.e. VI) and the test with best assumed DSe was considered the resolver test (i.e. RT-PCR). Concerns about conditional dependence among tests were expressed, but not addressed, by the authors.

The last approach for PGS procedures is referred to as the **study pseudo-gold standard** (SPGS), which groups any available information, including test results (or not), for the ad-hoc definition of a standard that is specific to the study (Dohoo et al., 2009). Clear and rational explanations are required to justify the criteria of classification. When the standard definition includes results from the candidate test, incorporation bias may be a concern. In their study, Nérette et al. (2008b) used a SPGS based on results from three

ISAV tests to define the infection status of samples. In this instance, to be deemed infected at least two of the three tests had to be positive. In addition to incorporation bias, the authors suspected an overestimation of DSe and an underestimation of DSp due to misclassification. Nonetheless, both CRS and SPGS approaches provided similar estimates of test characteristics (Nérette et al., 2008b). In an earlier study, McClure et al. (2005) used a SPGS based on clinical evidence and not on detection assays to evaluate four ISAV tests. Infection status was identified according to the history of the cage of origin of the sampled salmon. A fish was deemed non-infected if it was collected from a site where none of the cages experience an ISAV outbreak, whereas a fish was deemed infected if it came from a cage experiencing increased mortality that was officially declared infected by the regulatory authorities (McClure et al., 2005). Although fish from a cage with a negative ISAV history had considerable evidence that they were not infected, it is questionable that all fish from an outbreak cage were infected (i.e. 100% prevalence). However, because the data originally came from surveillance records, all the fish that were sampled from outbreak cages were assumed to be moribund or recently deceased, which likely increases the prevalence closer to 100% (McClure et al., 2004). The authors recognized the limitations and potential bias of this SPGS expecting some overestimation of DSp and DSe since studied fish represent the extremities of the gradient of infection stages (either "strong" non-infected or heavily infected) (McClure et al., 2005).

For all PGS procedures, it is assumed that every sample is retested. If this is not the case, verification bias is expected and adjustment procedures for partial verification should be considered.

When standards are not available or their performance estimates are not reliable, alternative approaches have been developed to estimate test DSe and DSp **without a reference standard** (Fig. 1.5).

The first approach assumed non-lethal or minorly invasive sampling techniques that enabled the investigator to repeatedly test the same individual (Schulzer et al., 1991). Although it was expected to increase the confidence of the true status of the animal, it requires that the infection status stay the same during the entire sampling period. With a slow progression over time, chronic disease might be more suitable for this type of design. In addition, **repeated samples** may be used in experimental infection challenges to investigate the performance of the tests during the progression of the infection. Without measurable markers to identify the beginning and the end of the infection, the status of each individual is still unknown at a specified time. In addition, dependence among test results from the same individual should be accounted for in the analysis to avoid point and variance estimation bias (Greiner & Gardner, 2000b). Recently, an analytical approach was developed to evaluate tests using repeated samples when no gold standard is available (Engel et al., submitted).

Another alternative when no reliable standard is available employs the opinion of diagnostic experts and is referred to as **Delphi estimation**. This method does not require any sample collection or testing. A detailed questionnaire is submitted to a selected panel of experts requesting their opinion on the performances of defined tests under specified conditions. The first round of answers is then summarized before resubmitting the same questions a second time with the addition of the summary of the first set of answers. The

Delphi method was previously used to assess the accuracy of screening tests for two aquatic viral diseases (Infection Pancreatic Necrosis Virus and Infectious Hematopoietic Necrosis) (Bruneau et al., 1999). Although useful when no other information is available, this estimation method relies on the subjective opinion of the panel (i.e. information bias). For instance, diagnostic experts are often assumed to be test developers or operators from laboratories who may not have clinical, epidemiological or population perspectives. Due to the important subjectivity and reduced precision of this method, Delphi estimation is only used when evidence-based approaches are not available.

The last method to evaluate diagnostic tests without reference information was developed by Hui & Walter (1980) using **latent class model** (LCM). This method relies on an alternative statistical approach that does not consider tested samples separately (i.e. binary outcome: D+ or D-) but as part of a larger cluster or population (i.e. probability to be infected/disease based on prevalence). Therefore, study specimens are required to come from at least two separate populations with different prevalences, and to be tested by at least two tests. Since the existence of reliable reference standard tests is a recurring challenge, latent class models are progressively becoming the preferred method for diagnostic test evaluation. Guidelines adapted to the veterinary context were developed to facilitate the application of this complex procedure (Mintiens et al., submitted). However, LCM requires several populations to estimate one set of DSe and DSp and these are closely associated with the mixture distribution of covariate factors from a specific population. Therefore, pooled estimates of DSe/DSp are expected to be biased (i.e. case mix bias, Begg, 1987). In addition, LCM uses the information from the candidate to

estimate DSe/DSp, and therefore some degree of incorporation bias is suspected. The design and requirements for this estimation procedure are explained in section 1.5.2.2.2.

Evaluation parameters and standards for comparisons are of primary concern, and the identification and collection of appropriate samples in sufficient number represent a further challenge for diagnostic evaluations.

1.5.1.2 Sampling considerations

1.5.1.2.1 Number of samples

Different approaches have been described to calculate the sample size according to the objectives of the study (Obuchowski, 1998; Branscum et al., 2007), including:

1. estimating the DSe and the DSp of a candidate test,
2. determining, with a specified level of confidence, if the DSe and DSp of a candidate test is above a fixed value,
3. comparing the DSe and DSp of a candidate test to a fixed value,
4. determining if the DSe and DSp of two tests are equivalent, and
5. determining if a candidate test is superior (either or both DSe and DSp) to a reference test.

When a GS is available, the D+ (cases) and the D- are sampled separately in a sampling design referred to as *case-control* (see section 1.5.1.2.2). Obuchowski (1998)

reviewed the different calculations that could be used when cases and controls are sampled separately. When a GS is not available, the D+ and D- are sampled together in a sampling design referred to as *cross-sectional* (see section 1.5.1.2.2). Branscum et al. (2007) reviewed the different computations that can be implemented in a Bayesian framework when the true health status is unknown.

For the evaluation of DSe and DSp of a candidate test (1), tentative standards of sample sizes have been used without clear justifications. Some previous examples of sample size recommendations have been quite unrealistically high, such as 300 D+ and 1000 to 5000 D- sampled individuals suggested by Jacobson (1998). Conventionally, it seemed generally accepted that sample size calculations for estimates of DSe/DSp follow the normal approximation of the binomial distribution as previously described for a single proportion (given equation Eq. (30), section 1.4.1.2.1) (Jacobson, 1998; Greiner & Gardner, 2000b; OIE, 2007; Dohoo et al., 2009). Increasing the sample size to ensure that coverage of the exact binomial confidence interval is respected has also been recommended (Greiner & Gardner, 2000b). However, modern software can easily perform exact binomial calculations (Flahaut et al., 2005). When the health status is unknown, Branscum et al. (2008) described calculation procedures for one test in one population.

As a preliminary evaluation, one might want to ensure that the DSe or DSp is at least above a certain value (2). For instance, the investigator may want to confirm that the assay meets minimum requirements before proceeding to more intensive testing. Using the case-control sampling approach, one option would be to use the sample size calculation to demonstrate freedom of disease (Dohoo et al., 2009). Assuming that the

population of D+ (or D-) is infinite, a minimum of five randomly selected samples would be necessary to test that the DSe (or DSp) is above 50% with 95% confidence (type I error, $\alpha = 5\%$). If one sample tested negative (or positive) the DSe (or DSp) would be assumed to be below 50%. For 99% level of confidence, 7 samples per population would be necessary.

Towards a more precise evaluation, the DSe or DSp of the candidate test can be compared to a fixed value (3). Reviewed by Obuchowski (1998) and illustrated in Flahaut et al. (2005), this approach adapts the sample size calculation formulae to compare two proportions (Dohoo et al., 2009). For an expected DSe, the number of D+ samples required to estimate, with a probability of $1 - \beta$, that DSe is at least higher than a defined value DSe_{min} (i.e. lower bound of the confidence interval) with a minimum confidence level of $1 - \alpha$ (one sided) is:

$$n = [Z_{1-\alpha} \sqrt{((DSe-DSe_{min})\ (1-DSe+DSe_{min}))} + Z_{1-\beta} \sqrt{(DSe(1-DSe))}]^2 / (DSe-DSe_{min})^2 \quad (38)$$

Conservatively, the fixed value is set at 50%. The number of D- samples can be computed separately or Flahaut et al. (2005) recommended calculating the number of D- (n_{NI}) from the number of infected (n_I), based on the prevalence of the study population (Pr):

$$n_{NI} = n_I * (1-Pr) / Pr \quad (39)$$

The prevalence is frequently expected to be lower than 50%, resulting in a requirement for more D- than D+ samples. Pre-computed sample sizes for several sets of proportions are summarized in Appendix 4.

When the true health status is known, more complex sample size calculations are available to compare two tests in paired or unpaired designs (i.e. when the two assays test the same samples or not) to evaluate equivalence (4) or define if one is superior to the other (5) (Obuchowski, 1998). For diagnostic evaluation in absence of a standard test (i.e. latent class models), Georgiadis et al. (2005) developed a spreadsheet-based program that, for a set level of precision, computes the required sample size for LCM analysis involving 2 populations and 2 tests. For more complex situations where conditional dependence exists between tests and/or more than 2 populations are sampled, Branscum et al. (2007) described a Bayesian approach to compare assays or test equivalency.

In addition, when DSe and DSp are not the parameters of interest, sample size calculations have been described for likelihood ratios (LR) and area under the Receiver Operating Characteristic (ROC) curve (continuous outcome test) (Simel et al., 1991; Obuchowski, 1998).

While an appropriate sample size calculation can be difficult to implement, the representativeness of the samples is also a primary concern to properly evaluate the true natural variation of the test performance in the targeted population.

1.5.1.2.2 Origin of samples

The case of one target population is addressed in this section, as the reviewed principles can be repeated with additional populations when more than one population is

included (e.g. LCM). Similar to agreement estimation, it is important that samples collected for trueness estimation are as similar as possible to those that are submitted during routine practice (i.e. representative of the targeted population). The nature, homogeneity and stability of diagnostic study specimens were discussed in detail previously (see section 1.3.1.2.2). Also, the direction of the data collection can be defined as retrospective, ambispective, or prospective relative to the period when the information was gathered about the health status and/or candidate test results (Knottnerus & Muris, 2003). Several sampling designs were described in past diagnostic evaluation studies with various degrees of validity. We group them in four categories: case-control, cross-sectional, cohort, and clinical field trial.

Case-control sampling. This design refers to study where D+ and D- are sampled separately. The disease status of the specimen is known at the time of the collection assuming that the method to determine the status is perfect (gold standard). This approach is more intuitive to estimate DSe/DSp and convenient when D+ individuals are rare (i.e. low prevalence) (Greiner & Gardner, 2000b). However, to increase the certainty about the status of the animal, either highly D+ (advanced clinical manifestation) or highly D- (never exposed) individuals are sampled which only represent the extreme ends of the natural gradient of infection. Referred to as spectrum bias (a type of selection bias), this sampling design is expected to overestimate both DSe and DSp (Ransohoff & Feinstein, 1978). Three commonly used approaches for case-control sampling that avoid the utilisation of a controversial GS test are outlined below.

Disease free population- To estimate DSp, the investigator may sample only animals from a disease-free population. Any specimens that test positive are thereafter deemed false, and DSp is directly estimated as the proportion of samples that test negative. A population is considered free with an acceptable level of certainty if no historical evidence of disease was documented and/or an active targeted surveillance program to confirm no disease present. To be valid, the disease-free population should be representative of populations that may be tested and fish should be collected randomly.

Experimental challenge- Although considered more representative than a "spike" sample, specimens from experimentally challenged animals are rarely representative of a natural infection episode for which the test will be employed. Depending on the infective dose, pathogen source, origin of the animal (e.g. certified specific pathogen free), or environmental conditions, the induced disease can be over- or under-expressed. Albeit convenient when naturally infected animals are not accessible, OIE recommends not using this sample source for evaluating diagnostic tests (OIE, 2009b).

Archived specimens- Archived samples from diagnostic laboratories are often used as convenient inputs for diagnostic test evaluations (i.e. retrospective study). The true status of the cases is identified according to extensive historical information and clinical evidence. Although possible to select obvious cases, the selection from the archive should be conducted in a formally random manner. Except for laboratories with large caseloads (i.e. many samples from the same population over a short duration), this type of sampling is not expected to be representative of the infection spectrum (Greiner & Gardner, 2000b). Indeed, the infection severity spectrum of submitted cases may change

over time and across origins, and the subsequent combinations of specimens may result in an unrealistic profile. Furthermore, depending on the type of samples, the stability over time of the specimen should be validated before being included in the study.

Cross-sectional sampling- This design refers to studies where D+ and D- individuals are sampled together. Specimens are sampled randomly from a specified population similar to a prevalence or survey type study. This sampling design requires that the condition of interest (infection or disease) is common in the targeted population. Depending on the type of standard, the samples are thereafter tested simultaneously by the different assays involved in the study. This approach is expected to increase the likelihood that the study population will be represented properly.

Random sampling can be approached in different ways according to the presence of a sampling frame and the structure of the population (Dohoo et al., 2009): *simple random sampling* (sampling frame available), *systematic random sampling* (repeated collection in a logical manner), *stratified random sampling* (collection proportional to covariates expected to influence DSe/DSp), *clustered sampling* (collection of all individuals from randomly selected clusters), and *multistage sampling* (collection of a few randomly selected individual from randomly selected clusters). Under aquaculture settings, techniques to sample fish in a representative manner (i.e. randomly) may depend on the infection of interest and fish production methods. Spread into a three dimensional environment, capture of healthy fish is not convenient and the utility of various field methods have been reviewed elsewhere (Hammell, 1992; des Clers, 1994; Thorburn, 1999; Cameron, 2002; OIE, 2009b). Furthermore, Van den Bruel et al. (2006) discussed

that to be completely validated, a test evaluation study should be repeated on independent but similar populations to ensure external validity. Nonetheless, as discussed with retrospective sampling (e.g. archived specimens), infection patterns evolve and change over time which supports the concept of a test performance monitoring program over time to address concerns regarding single time point estimates.

Cohort sampling- A cohort identifies a group of individuals that share a common character (Dohoo et al., 2009). In this sampling design, animals are selected based on a common symptom and followed through time for an outcome to occur. The data collection is prospective and targets the estimation of the prognostic value of the candidate test (Lijmer & Bossuyt, 2009). This type of approach is only applicable with non-lethal sampling or when infected animals have a distinct disease with clinical manifestations (e.g. death). This design has very specific and limited applications (Hui & Zhou, 1998).

Clinical field trial- In this type of study, the health status of the individual is of secondary importance. Only the candidate test result is known and the outcome of interest is either clinical or economic effects. The design is prospective and targets the evaluation of the candidate test's predictive values for individuals that show minor or non-specific clinical evidence (Lijmer & Bossuyt, 2009). Based on the initial test result, refined treatment strategies are applied to tested individuals while non-tested individuals receive a default treatment. The usefulness of the test is then evaluated based on the animal outcome and/or treatment cost savings. Although applied to clinical settings, this type of

study selects individuals based on preliminary clinical signs to ensure a substantial proportion of D+. The evaluation of the test will therefore only be valid for similar conditions of utilization.

For veterinary applications, a test can be intended for a wide range of purposes and the evalutions will require a wide range of designs to reflect this. The sampling scheme with minimal bias is usually a cross-sectional sampling of the targeted or similar population. Estimation procedures that require samples from more than one population (e.g. LCM) may experience difficulties avoiding case mix bias (Begg, 1987).

1.5.1.3 Other associated bias

For diagnostic trueness studies, some precaution is required when handling or testing samples to avoid unknown bias. It is necessary to blind the test operator(s) using coded samples to avoid *review bias* (Greiner & Gardner, 2000b). Review bias refers to the potential change in test results when the operator has additional information (e.g. clinical sign) associated with a specimen. Ransohoff & Feinstein (1978) furthermore differentiated two types of review bias: *diagnostic-review bias* is referring to a systematic test error arising when sample-associated information is provided before the final diagnostic result is available; and *test-review bias* is a problem when the operator is informed of previous test results. In practice, review bias due to preliminary clinical information tends to over-estimate test performances (Loy & Irwig, 2004), but

diagnostic-review bias can also under-estimate performances (Ransohoff & Feinstein, 1978).

1.5.2 Analysis methodology

1.5.2.1 Trueness parameters (DSe, DSp, Ef, J, DOR)

Direct estimation methods for trueness parameters when a gold standard exists (i.e. true health status known) are discussed in this section. A "fourfold" or 2X2 table can be generated from the test and health status information (Feinstein, 1975):

	Positive Test	*Negative Test*	
Infected/Diseased	*a*	*b*	*a + b*
Non-infected/Non-diseased	*c*	*d*	*c + d*
	a + c	*b + d*	*n = a + b + c + d*

Thereafter, DSe and DSp can be estimated as follows:

$$DSe = a / (a + b) \quad (40)$$

$$DSp = d / (c + d) \quad (41)$$

Similar to SE_{Pa}, the standard errors of DSe and DSp (SE_{DSe} & SE_{DSp}) arae calculated as for any other proportion ($SE_{DSe} = \sqrt{(DSe*(1-DSe)/ (a + b))}$; $SE_{DSp} = \sqrt{(DSp*(1-DSp)/ (c + d))}$) with confidence intervals calculated using normal approximation, exact binomial, bootstrap or Jackknife approaches. The program Testview is one of many utilities that can be used to simultaneously estimate several validation criteria (Gardner & Holmes, 1993).

Ef can also be estimated from the 2X2 table or from DSe/DSp:

$$Ef = (a + d) \ / \ n \ = \ P*DSe + (1-P)*DSp \qquad (42)$$

Standard error and confidence intervals of Ef follow the same estimation principles for proportion statistics.

Youden's index (J) is expressed as the average of the "successes" (S) of the test in the D+ group and the D- group ($(S_{diseased} + S_{non\text{-}diseased})/2$). Success is calculated as the proportion of correctly classified individuals minus the proportion of incorrectly classified individuals in each disease group. Using the 2X2 table notation, the success of the test in the diseased group ($S_{diseased}$) is computed as $(a – b) / (a + b)$; and the success of the test in the non-diseased group ($S_{non\text{-}diseased}$) is computed as $(d – c) / (c + d)$. The index J can be directly estimated using the 2X2 table or DSe/DSp:

$$J = (ad - bc)/ [(a + b)(c + d)] = DSe + DSp - 1 \qquad (43)$$

Standard error of J (SE_J) can also be computed ($SE_J = \sqrt{((ab /(a+b)^3) + (cd/(c+d)^3))}$) and the confidence interval is approximated using a normal distribution, provided there are more than 20 individuals in each infection/disease group (Youden, 1950). To be useful, it is expected that the test will, on average, more often correctly classify samples and J will therefore range between 0 and 1 (positive discrimination). When J is negative, samples are more often misclassified (negative discrimination) meaning that there is more chance to be correct by inversing the test results. When there is no evidence of J being different

from 0, the test has no discriminatory power and is worthless. The test is perfect when J is equal to 1.

DOR can be estimated using the 2X2 table, DSe/DSp, PPV/NPV, or LR^+/LR^- in the following manner:

$$DOR = ad / cb = [DSe/(1-DSe)]/[(1-DSp)/DSp]$$
$$= [(1-PPV)/PPV]/[(1-NPV)/NPV] = LR^+/LR^- \quad (44)$$

Standard error for DOR (SE_{DOR}) is calculated as any other odds ratio ($SE_{DOR} = \sqrt{(1/a + 1/b + 1/c + 1/d)}$), and the confidence interval is computed using normal approximation when DOR is not close to 0 (Glas et al., 2003). When the DOR is significantly > 1, individuals have a greater chance to be correctly classified using the test. The greater DOR is above 1, the better the discriminatory accuracy of the test. When the DOR is significantly < 1, tested individuals have more chance to be misclassified (negative discrimination). With no evidence of DOR differing from 1, the test has no discriminatory capability.

1.5.2.2 Estimation procedure without a gold standard

1.5.2.2.1 Using an imperfect standard

Even if a gold standard is unavailable, it is possible to estimate DSe/DSp when the operating characteristics of an imperfect reference standard are known and the two

tests are conditionally independent (Dohoo et al., 2009). Based on the following contingency table:

	Candidate Test		
Reference test	*Positive Test*	*Negative Test*	
Positive Test	*a*	*b*	*a + b*
Negative Test	*c*	*d*	*c + d*
	a + c	*b + d*	*n = a + b + c + d*

the candidate test characteristics can be calculated as follows (Enøe et al., 2000):

$$DSe_{cand} = ((a+c)\, DSp_{ref} - c) / (n\, DSp_{ref} - (c+d)) \quad (45)$$

$$DSp_{cand} = ((b+d)\, DSe_{ref} - b) / (n\, DSe_{ref} - (a+b)) \quad (46)$$

Standard error and confidence interval estimation are described elsewhere (Buck & Gart, 1966; Gart & Buck, 1966).This calculation is only possible because the data and the parameters are both three-dimensional (3 equations for 3 unknowns: DSe, DSp and prevalence). When the tests are conditionally dependent, covariate factors must also be estimated and the number parameters to estimate becomes greater than the data dimension (i.e. 3 equations for 5 unknowns: DSe, DSp, prevalence, cov+ and cov-).

However, the previous evaluation of DSe and DSp of the reference standard might not be reliable or valid. Therefore, the estimation of operating characteristics of the candidate test may be inherently biased because of the originally biased estimates for the reference standard. When no standard tests are available or reliable, statistical methods have been developed to evaluate more than 2 tests on the same specimen from at least 2 populations.

1.5.2.2.2 Without reference standard

When no standard is available, Hui & Walter (1980) adapted a latent class model (LCM) estimation that does not require information about the true status. However, with no previous information, this procedure requires a minimum of 2 tests run on the same sample from at least 2 populations with different prevalences. The mechanics of the LCM procedure based on 2 tests and 2 populations is described in Appendix 5. Three assumptions are necessary when using LCM for diagnostic evaluations:

(i) *At least two dichotomous assays should be applied to the same individuals from at least two populations* (with different assumed prevalences). This assumption is a study design requirement to ensure that the model is *identifiable* when no prior information is known about the tests or the source populations. In statistics, the "identifiability" refers to the ability of a model to be developed based on all the unknown parameters (DSe, DSp and prevalences) and the information available from the data (i.e. degrees of freedom, *df*). For instance, the number of parameters in the 2 tests-2 populations model equals the data *df* (i.e. 6). However, no residual *df* are available to evaluate the validity and goodness-of-fit of the model, and therefore more than two tests and/or populations are preferred (Dohoo et al., 2009). Evaluations of LCM identifiability applied to diagnostic tests have been discussed elsewhere (Jones et al., 2009). LCM models can also be run with either fewer tests or fewer populations when prior information about the prevalence or test properties exist (Dendukuri & Joseph, 2001). However, in some situations, the prior information may have more weight on the

estimation than the data itself (Neath & Samaniego, 1997). In these models, the number of tests should be increased or the available populations stratified (Toft et al., 2005). For instance, for a single sampled population, a minimum of 4 tests are necessary (Dendukuri and Joseph, 2001).

(ii) *The tests must be conditionally independent given the infection/disease status*. Test independence conditional on the health status may not be realistic when several tests are used (Brenner, 1996) and can result in incorrect estimation due to information bias (Torrance-Rynard & Walter, 1997). Conditional dependence was explained in section 1.2.3.9, and the methods developed to account for conditional dependence in LCM are reviewed elsewhere (Hui and Zhou, 1998; Toft et al., 2005; Branscum et al., 2005; Dendukuri et al., 2009).

(iii) *The operating characteristics of the tests must be constant across populations*. Although generally assumed, constant DSe and DSp for various prevalences is actually rare. DSe is expected to increase with prevalence while DSp is expected to decrease (Greiner & Gardner, 2000b). When the assumption of constant DSe and DSp is not valid, the LCM may result in estimates that pool DSe/DSp from the different populations (Toft et al., 2005). The impact on LCM of varying DSe and DSp was investigated and discussed elsewhere (Johnson et al., 2009). This assumption is rarely verified in LCM. DSe and DSp estimations in the different studied populations can be done using pseudo-gold standard procedures (Nérette et al., 2008b). However, no

modelling procedure has been developed to relax this assumption without increasing substantially the number of parameters, which can jeopardize the model identifiability.

Two different estimation procedures exist to fit LCM and they include: maximization of the likelihood function (maximum likelihood, ML) and Bayesian estimations. The ML estimates can be obtained using an Expectation Maximization (EM) algorithm and the corresponding standard errors using the Newton-Raphson estimation (Dohoo et al., 2009). Confidence intervals are usually obtained by bootstrapping. The "Tests in the Absence of a Gold Standard" (TAGS) software was developed to facilitate the utilization of this procedure using either an "R" interface or directly on the website http://www.epi.ucdavis.edu/diagnostictests/QUERY.HTM (Pouillot et al., 2002). However, ML estimation is not particularly flexible and cannot incorporate prior information about parameters. Furthermore, it does not perform properly when the sample size is small or when cells from contingency tables include nil or small values (Dohoo et al., 2009).

The Bayesian framework accommodates the modification of the code to include prior information or additional parameters (Hui & Zhou, 1998). Usually Bayesian LCM is run in the WinBugs software (Spiegelhalter et al., 2003) and codes for various estimation scenarios are available online (http://www.epi.ucdavis.edu/diagnostictests/software.html#DiagnosticTestSeSp). Associated with Bayesian estimation, post-estimation guidelines to assess Markov Chain Monte Caro convergence in the specific context of diagnostic test evaluation has been outlined elsewhere (Toft et al., 2007). A friendly interface was recently developed to use

LCM with two or more classes from a single population (http://www.nandinidendukuri.com) and a version for several populations will soon be available (Nandini Dendukuri, pers. com.). As multiple Bayesian outputs (i.e. mean, median or mode) are obtained, which one is reported should be clearly stated or all 3 should be reported, to ease subsequent interpretations and applications.

In the specific context of ISAV diagnostic test evaluations, LCM models have used both ML (Nérette et al., 2005a; Gustafson et al., 2008) or Bayesian estimations (Nérette et al., 2008b). Although, the assumption of conditional independence among tests was evaluated and accounted for in the estimation procedure (Nérette et al., 2008b), variation in DSe and DSp across the sampled populations was acknowledged but not addressed.

1.6 Thesis Rationale: Factors affecting diagnostic accuracy

In 1947, Yerushalmy understood that the trueness of a diagnostic test could not be summarized as a single parameter (i.e. Ef) that would strongly vary with prevalence of D+ in the tested population (Alberg et al., 2004). Therefore, he stratified the proportion of correctly classified specimens within D+ and D- individuals, defined as DSe and DSp respectively. The concept of relative performance was not considered for agreement evaluation. Similarly, repeatability and reproducibility are prevalence-weighted averages of agreement within D+ animals and agreement within D- animals. When status-specific estimates are substantially different from each other or when the assumed prevalence of the targeted population deviates from 50%, the overall agreement varies with prevalence.

Albeit agreement studies are focused on test comparability or discrepancy, estimates of agreement relative to the health status seem more relevant to interpret and predict agreement variation in various populations.

The health status is certainly not the only factor that influences the classification performance of a detection method. Contrary to common understanding, DSe and DSp may not be constant parameters across populations. Greiner & Gardner (2000b) readdressed DSe and DSp as population parameters that vary within and between populations according to the distribution of biological factors that influence the disease biology. For instance, age was found to be positively correlated with DSe of bovine trypanosomosis serology (Greiner et al., 1997).The validity of diagnostic accuracy studies is therefore questionable when the spectrum of covariate factors differs between the study population and target population(s). OIE addresses this issue by requiring that assay evaluation studies use a large number of realistic and representative samples. Nonetheless, the apparent poor stability of reported DSe/DSp warrants more consideration to these parameters (Moons & Harrell, 2003). Conversely, the widely adopted Standards for Reporting of Diagnostic Accuracy (STARD) recommended to keep using DSe/DSp and to address their relativity by reporting various population-specific estimates (Bossuyt et al., 2003). Nonetheless, this approach does not allow for a valid extrapolation of test characteristics to a new, unstudied population. Obtained by stratification or modelling, covariate-specific estimates of DSe and DSp are more relevant to investigate the impact of specified factors on test accuracy (Shapiro, 1999, Janes & Pepe, 2008; Bachmann et al., 2009). Finally, according to the assumed mixture

distribution of covariates in the population, DSe and DSp can be predicted across various populations by simple weighted averages (Björk et al., 2009).

The most influential factor impacting DSe is most likely the degree of disease severity. Based on histology or morphology, DSe is expected to be greater for large and diffuse lesions than for small and focal lesions (Begg, 1987). Moreover, the spectrum of virulence and immunogenicity of an infectious disease is assumed to vary across populations according to the prevalence (Brenner & Gefeller, 1997). With an increasing prevalence, the proportion of severe stages is expected to increase among infected animals (Greiner & Gardner, 2000b). Therefore, constant DSe across prevalences is unlikely and is expected to increase with prevalence. Furthermore, it is reasonable to think that test performance in D- individuals (DSp) is dependent on the pressure of cross-contamination from infected samples or/and positive controls. As the proportion of infected animals in the sampled population (prevalence) and the tested sample pool become larger, the pressure and chance of cross-contamination likely becomes greater. Therefore, DSp is expected to decrease with increased prevalence. Referred to as "mix case" (Begg, 1987), misrepresentation of infection stage may result in biased pooled estimations for procedures like LCM that assume constant DSe and DSp across prevalence populations (Toft et al., 2005).

Finally, detection assays that measure a continuous biological marker involve an additional level of dependence. The two distributions corresponding to D+ and D- group measurements tend to partially overlap each other. The degree of overlapping depends on various factors that characterize the target population (e.g. spectrum of severity) (Greiner & Gardner, 2000b). According to the mixture distribution and the intended purpose of the

test, a cut-point value is selected to discriminate D+ and D- individuals. For the same test and with the same objective, the cut-point (and associated DSe/DSp) is expected to change across populations. Therefore, the selection of a fixed and standard cut-point in continuous outcome tests may not be appropriate.

1.7 Thesis objectives

The main objective of this research was to develop methods that account for factors that are likely to affect the estimation of test accuracy fro ISAV. Addressed in four separate research chapters, the specific objectives were:

1. to extend the basic description of test agreement by adding visual tools to summarize and illustrate the relative agreement among pairs of tests or runs.

2. to estimate agreement related to the true health status with subsequent predictions of the overall agreement in other populations, for a specified prevalence.

3. to extend the conventional latent class model by adding a third class (subclass of infected individuals), and to achieve constant DSe estimates based on the stage of infection (i.e. relax the assumption of constant DSe).

4. to adapt the selection of cutpoints for the particular case of real-time nucleic acid amplification assays based on the intended purpose and the target population.

1.8 References

Agresti, A., 2002. Categorical data analysis. New York: Wiley.

Akobeng, A.K., 2007a. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr. 96, 338-341.

Akobeng, A.K., 2007b. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. Acta Paediatr. 96, 487-491.

Alberg, A.J., Park, J.W., Hager, B.W., Brock, M.V., Diener-West, M., 2004. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J. Gen. Intern. Med. 19, 460-465.

Alonzo, T.A., Pepe, M.S., 1999. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Statist. Med. 18 , 2987-3003.

Anonymous, 2000. ISA hits the Faroes. Fish Farming International. 27, 47.

Aspehaug, V., Mikalsen, A.B., Snow, M., Biering, E., Villoing, S. 2005. Characterization of the infectious salmon anemia virus fusion protein. J. Virol. 79, 12544-12553.

Bachmann, L.M., ter Riet, G., Weber, W.E., Kessels, A.G., 2009. Multivariable adjustments counteract spectrum and test review bias in accuracy studies. J. Clin. Epidemiol. 62,357-361.

Begg, C.B., 1987. Biases in the assessment of diagnostic tests. Stat. Med. 6, 411-423.

Begg, C.B., Greenes, A., 1983. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39, 207–215.

Bertrand, P., Bénichou, J., Grenier, P., Chastang, C., 2005. Hui and Walter's latent-class reference-free approach may be more useful than diagnostic performance in assessing agreement. J. Clin. Epidemiol. 58, 689-701.

Björk, J., Grubb, A., Nyman, U., (2009). Variability in diagnostic accuracy can be estimated using simple population weighting. J. Clin. Epidemiol. 62, 54-7.

Blake, S., Bouchard, D., Keleher, W., Optiz, M., Nicholson, B.L., 1999. Genomic relationship of the North American isolate of infectious salmon anaemia virus ISAV) to theNorwegian strain of ISAV. Dis. Aquat. Organ. 35, 139-144.

Bols, N.C., Ganassin, R.C., Tom, D.J., Lee, L.E., 1994. Growth of fish cell lines in glutamine-free media. Cytotechnology 16, 159-166.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C.W., 2003. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin. Radiol. 58, 575-580.

Bouchard, D.A., Brockway, K., Giray, C., Keleher, W., Merrill, P.L., 2001. First report of infectious salmon anaemia (ISA) in the United States. Bull. Eur. Assoc. Fish. Pathol. 21, 86-88.

Bouchard, D., Keleher, W., Opitz, H.M., Blake, S., Edwards, K.C., Nicholson, B.L., 1999. Isolation of infectious salmon anaemia virus (ISAV) from Atlantic salmon in New Brunswick, Canada. Dis. Aquat. Organ. 35, 131-137.

Branscum, A.J., Gardner, I.A., Johnson, W.O., 2005. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling**.** Prev. Vet. Med. 68, 145-163.

Branscum, A.J., Johnson, W.O., Gardner, I.A., 2007. Sample size calculations for studies designed to evaluate diagnostic test accuracy. J. Agri. Biol. Environ. Statist. 12, 112-127.

Brenner, H., 1996. How independent are multiple "independent" diagnostic classifications? Stat. Med. 15, 1377-1386.

Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat. Med. 16, 981-991.

Brown, L.L., Sperker, S.A., Clouthier, S., Thornton, J.C., 2000. Development of a vaccine against infectious salmon anaemia virus (ISAV). Bull. Aquat. Ass. Canada 100, 4-7.

Bruneau, N.N., Thorburn, M.A., Stevenson, R.M.W., 1999. Use of the Delphi panel method to assess expert perception of the accuracy of screening test systems for infectious pancreatic necrosis virus and infectious hematopoietic necrosis virus. J. Aquat. Anim. Health 11, 139-147.

Buck, A.A., Gart, J.J., 1966. Comparison of a screening test and a reference test in epidemiologic studies. I. Indices of agreement and their relation to prevalence. Am. J. Epidemiol. 83, 586-592.

Byrne, P.J., MacPhee, D.D., Ostland, V.E., Johnson, G., Ferguson, H.W., 1998. Haemorrhagic kidney syndrome of Atlantic salmon, *Salmo salar* L. J. Fish Dis. 21, 81-91.

Byrt, T., Bishop, J., Carlin, J.B., 1993. Bias, prevalence and kappa. J. Clin. Epidemiol. 46, 423-429.

Cameron, A.R., Baldock, F.C., 1998a. A new probability formula for surveys to substantiate freedom from disease. Prev. Vet. Med. 34, 1-17.

Cameron, A.R., Baldock, F.C., 1998b. Two-stage sampling in surveys to substantiate freedom from disease. Prev. Vet. Med. 34, 19-30.

Cameron, A.R., 2002. Survey Toolbox for Aquatic Animal Diseases: A Practical Manual and Software Package. CSIRO Publishing.pp. 373.

Cannon, R.M., Roe, R.T., 1982. Livestock disease surveys: a field manual for veterinarians. Australian Bureau of Animal Health, Canberra.

Christensen, J., Gardner, I.A., 2000. Herd-level interpretation of test results for epidemiologic studies of animal diseases. Prev. Vet. Med. 45, 83-106.

Christiansen, D.H., Ostergaard, P.S., 2008. Prevalence and genetics of low pathogenic infectious salmon anemia virus in farmed Atlantic salmon (*Salmo salar* L) in the Faroes. American Fisheries Society, Fish health section - Annual meeting, July 9-12, 2008, Charlottetown , Canada.

Cipriano, R.C., 2002. Infectious Salmon Anemia Virus. Fish Disease Leaflet # 85 at the United States Geological Survey, National Fish Health Research Laboratory, WV 25430, U.S.A. http://www.lsc.usgs.gov/fhb/leaflets/FHB85.pdf. 11 pp.

Cipriano, R.C., 2009. Antibody against infectious salmon anaemia virus among feral Atlantic salmon (Salmo salar). ICES J. Mar. Sci. 66, 865-870.

Cleophas, T.J., Droogendijk, J., van Ouwerkerk, B.M., 2008. Validating diagnostic tests, correct and incorrect methods, new developments. Curr. Clin. Pharmacol. 3, 70-76.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 37-46.

Cowling, D.W., Gardner, I.A., Johnson, W.O., 1999. Comparison of methods for estimation of individual-level prevalence based on pooled samples. Prev. Vet. Med. 39, 211-225.

Crowther, J.R., Unger, H., Viljoen, G.J., 2006. Aspect of kit validation for test used for the diagnosis and surveillance of livestock diseases: producer and end-user responsibilities. Rev, Sci. Tech. Off. Int. Epiz. 25, 913-935.

Dannevig, B.H., Falk, K., 1994. Atlantic salmon (*Salmo salar* L.) develop infectious salmon anaemia (ISA) after inoculation with in vitro infected leukocytes. J. Fish Dis. 17, 183-187.

Dannevig, B.H., Falk, K., Krogsrud, J., 1993. Leucocytes from Atlantic salmon (*Salmo salar* L) experimentally infected with infectious salmon anaemia (ISA) exhibit an impaired response in mitogens. J. Fish Dis. 16, 351-359.

Dannevig, B.H., Falk, K., Namork, E., 1995a. Isolation of causal virus of infectious salmon anaemia (ISA) in a long term cell line from Atlantic salmon head kidney. J. Gen. Virol. 76, 1353-1359.

Dannevig, B.H., Falk, K., Press, C.M., 1995b. Propagation of infectious salmon anemia (ISA) virus in cell culture. Vet. Res. 26, 438.442.

Dannevig, B.H., Brudeseth, B.E., Gjoen, T., Rode, M., Wergeland, H.I., Evensen, O., Press, C.M., 1997. Characterization of a long-term cell line (SHK-1) developed from the head kidney of Atlantic salmon (*Salmo salar* L.). Fish Shellfish Immunol. 7: 213–226.

Deeks, J.J., Altman, D.G., 2004. Diagnostic tests 4: likelihood ratios. BMJ 329, 168-169.

Dendukuri, N., Joseph, L., 2001. Bayesian Approaches to Modeling the Conditional Dependence Between Multiple Diagnostic Tests. Biometrics 57:158-167.

Dendukuri, N., Hadgu, A., Wang, L., 2009. Modeling conditional dependence between diagnostic tests: A multiple latent variable model. Stat. Med. 28:441-461.

Des Clers, S., 1994. Sampling to detect infections and estimate prevalence in aquaculture, Pisces Press, Stirling, U.K.

Devold, M., Krossøy, B., Aspehaug, V., Nylund, A., 2000. Use of RT-PCR for diagnosis of infectious salmon anaemia virus (ISAV) in carrier sea trout *Salmo trutta* after experimental infection. Dis. Aquat. Organ. 40, 9-18.

Dohoo, I., Martin, W., Stryhn, H. (Eds.), 2009. Veterinary Epidemiologic Research. 2nd ed., AVC Inc., Charlottetown, Canada.

Donald, A.W., Gardner, I.A., Wiggins, A.D., 1994. Cut-off points for aggregate herd testing in the presence of disease clustering and correlation of test errors. Prev. Vet. Med. 19, 167-187.

Engel, B., Backer, J., Buist, W., Submitted. Evaluation of the Accuracy of Diagnostic Tests From Repeated Measurements Without a Gold Standard. J. Agri. Biol. Environ.Statist.

Enøe, C., Georgiadis, M.P.A., Johnson, W.O., 2000. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev. Vet. Med. 45, 61-81.

Evensen, O., Thorud, K.E., Olsen, Y.A., 1991. A morphological study of the gross and light microscopic lesions of infectious anaemia in Atlantic salmon (*Salmo salar*). Res.Vet. Sc. 51, 215-222.

Fagan, T.J., 1975. Letter: Nomogram for Bayes theorem. N. Engl. J. Med. 293, 257.

FAO, 2009. The state of world fisheries and aquaculture 2008. Part 1: World review of fisheries and aquaculture. Food and Agriculture Organization of the United Nations, Fisheries and Aquaculture Department, Rome. http://www.fao.org/docrep/011/i0250e/i0250e00.htm.

Falk, K., Dannevig, B.H., 1995. Demonstration of infectious salmon anaemia (ISA) viral antigens in cell cultures and tissue sections. Vet. Res. 26, 499–504.

Falk, K., Namork, E., Rimstad, E., Mjaaland, S., Dannevig, B.H., 1997. Characterization of infectious salmon anaemia virus, an orthomyxo-like virus isolated from Atlantic salmon (*Salmo salar* L.). Journal of Virology 71: 9016–9023.

Falk, K., Namork, E., Dannevig, B.H., 1998. Characterization and applications of a monoclonal antibody against infectious salmon anaemia virus. Dis. Aquat. Org. 34, 77–85.

FDA, 2007. Statistical guidance on reporting resuls from studies evaluating diagnostic tests. Guidance for industry and FDA staff. Document# 1620. http://www.fda.gov/cdrh/osb/guidance/1620.pdf. 39pp.

Feinstein, A.R., 1975. Clinical biostatistics. XXXIV. The other side of 'statistical significance': alpha, beta, delta, and the calculation of sample size. Clin. Pharmacol. Ther. 18, 491-505.

Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiol. 43, 543-549.

Flahaut, A., Cadilhac, M., Thomas, G., 2005. Sample size calculation should be performed for design accuracy in diagnostic test studies. J. Clin. Epidemiol. 58, 859-862.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical Methods for Rates and Proportions, 3rd Edition. Wiley, New York, USA.

Gardner, I.A., Holmes, J.C., 1993. Test view: a spreadsheet program for evaluation and interpretation of diagnostic tests. Prev. Vet. Med. 17, 9-18.

Gardner, I., Stryhn, H., Lind, P., Collins, M., 2000. Conditional dependence between tests affects the diagnosis and surveillance of animal disease. Prev. Vet. Med. 45, 107-122.

Gart, J. J., Buck, A.A., 1966. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for comparison of diagnostic tests. Am. J. Epidemiol. 83, 593-602.

Geisser, S., Johnson, W., 1992. Optimal administration of dual screening tests for detecting a characteristic with special reference to low prevalence diseases. Biometrics 48, 839-852.

Georgiadis, M.P., Johnson, W.O., Gardner, I.A., 2005. Sample size determination for estimation of the accuracy of two conditionally independent tests in the absence of a gold standard. Prev. Vet. Med. 71, 1-10.

Giray, C., Opitz, H.M., MacLean, S., Bouchard, D., 2005. Comparison of lethal versus non-lethal sample sources for the detection of infectious salmon anemia virus (ISAV). Dis. Aquat. Organ. 23, 181-185.

Glas, A.S., Lijmer, J.G., Prins, M.H., Bonsel, G.J., Bossuyt, P.M., 2003. The diagnostic odds ratio: a single indicator of test performance. J. Clin. Epidemiol. 56, 1129-1135.

Godoy, M.G., Aedo, A., Kibenge, M.J., Groman, D.B., Yason, C.V., Grothusen, H., Lisperguer, A., Calbucura, M., Avendaño, F., Imilán, M., Jarpa, M., Kibenge, F.S., 2008. First detection, isolation and molecular characterization of infectious salmon anaemia virus associated with clinical disease in farmed Atlantic salmon (Salmo salar) in Chile. BMC Vet. Res. 4, 28.

Grant, R., Smail, D.A., 2003. Comparative isolation of infectious salmon anemia virus (ISAV) from Scotland on TO, SHK-1, and CHSE-214 cells. Bull. Eur. Ass. Fish Pathol. 23, 80–85.

Gregory, A., 2002. Detection of infectious salmon anaemia virus (ISAV) by *in situ* hybridisation. Dis. Aquat. Organ. 50, 105-110.

Gregory, A., Munro, L.A., Snow, M., Urquhart, K.L., Murray, A.G., Raynard, R.S., 2009. An experimental investigation on aspects of infectious salmon anaemia virus (ISAV) infection dynamics in seawater Atlantic salmon, *Salmo salar* L. J. Fish Dis. 32, 481-9.

Greiner, M., Bhat, S.T., Patzelt, R.J., Kakaire, D., Schares, G., Dietz, E., Bohning, D., Zessin, K.H., Mehlitz, D., 1997. Impact of biological factors on the interpretation of bovine trypanosomiasis serology. Prev. Vet. Med. 30, 61-73.

Greiner, M., Gardner, I.A., 2000a. Application of diagnostic tests in veterinary epidemiologic studies. Prev. Vet. Med. 45, 43-59.

Greiner, M., Gardner, I.A., 2000b. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 42, 2-22.

Guggenmoos-Holzmann, I., 1996. The meaning of Kappa: concepts of reliability and validity revisited. J. Clin. Epidemiol. 49, 775-782.

Gustafson, L., Ellis, S., Bouchard, D., Robinson, T., Marenghi, F., Warg, J., Giray, C., 2008. Estimating diagnostic test accuracy for infectious salmon anaemia virus in Maine, USA. J. Fish Dis. 31, 117-125.

Gustafson, L., Ellis, S., Merrill, P., Robinson, T., MacPhee, D., 2005. Mortality rates predict apparent prevalence of infectious salmon anemia (ISA) at two infected Atlantic salmon farms in Maine. Bull. Eur. Ass. Fish Pathol. 25, 212-220.

Hadgu, A., 1996. The discrepancy in discrepant analysis. Lancet 348, 592-593.

Hammell, K.L., 1992. The relative bias in sampling estimates of production and disease parameters of caged Atlantic salmon. AVC MSc. Thesis. UPEI, Charlottetown, Canada.

Harel, O., Zhou, X.H., 2006. Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies. Statist. Med. 2715.

Hastings, T., Olivier, G., Cusack, R., Bricknell, I., Nylund, Å., Binde, M., 1999. Infectious salmon anaemia. Bull. Eur. Ass. Fish Pathol. 19,286-288.

Hiney, M.P., Smith, P.R., 1998. Validation of polymerase chain reaction-based techniques for proxy detection of bacterial fish pathogens: framework, problems and possible solutions for environmental applications. Aquaculture 162, 41-68.

Hjeltnes, B., Samuelsen, O.B., Svardal, A.M., 1992. Changes in plasma and liver glutathione levels in Atlantic salmon *Salmo salar* suffering from infectious salmon anemia (ISA). Dis. Aquat. Organ. 14, 31-33.

Hovland, T., Nylund, A., Watanabe, K., Endersen, C., 1994. Observation of infectious salmon anaemia virus in Atlantic salmon, *Salmo salar* L. J. Fish Dis. 17, 291-296.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.

Hui, S.L, Zhou, X.H., 1998. Evaluation of diagnostic tests without gold standards. Stat. Met. Med. Res. 7, 354-370.

ISO/IEC 17025:2005. General requirements for the competence of testing and calibration laboratories. 2nd Edition. International Standard. 28pp.

ISO International Standard 5725-1, 1994. Accuracy (trueness and precision) of measurement methods and results. Part 1: General principles and definition. International Organisation for Standardisation (ISO), ISO Central Secretariat, 1 rue de Varembé, Case Postale 56, CH - 1211, Geneva 20, Switzerland.

Jacobson, R.H., 1998. Validation of serological assays for diagnosis of infectious diseases. In Veterinary laboratories for infectious diseases (J.E. Pearson, ed.). Rev. Sci. Tech. Off. Int. Epiz. 17, 469-486.

Janes, H., Pepe, M.S., 2008. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. Am. J. Epidemiol. 168, 89-97.

Johnson, W.O., Su, C.L., Gardner, I.A., Christensen, R., 2004. Sample size calculations for surveys to substantiate freedom of populations from infectious agents. Biometrics. 60, 165-171

Johnson, W.O., Gardner, I.A., Metoyer, C.N., Branscum, A.J., 2009. On the interpretation of the test sensitivity in the two-test two-population problem: Assumptions matter. Prev. Vet. Med. 91, 116-121.

Jones, S.R.M., Mackinnon, A.M., Salonius, K., 1999. Vaccination of freshwater-reared Atlantic salmon reduces mortality associated with infectious salmon anaemia virus. Bull. Eur. Ass. Fish Pathol. 19, 98-101.

Jordan, D., 1996. Aggregate testing for the evaluation of Johne's disease herd status. Aust. Vet. J. 73, 16-19.

Jordan, D. McEwen, S.A., 1998. Herd-level test performance based on uncertain estimates of individual test performance, individual true prevalence and herd true prevalence. Prev. Vet. Med. 36, 187-209.

Joseph, T., Kibenge, M.T., Kibenge, F.S., 2003. Antibody-mediated growth of infectious salmon anaemia virus in macrophage-like fish cell lines. J. Gen. Virol. 84, 1701-1710.

Kawaoka, Y., Cox, N.J., Haller, O., Hongo, S., Kaverin, N., Klenk, H.-D., Lamb, R.A., McCauley, J., Palese, P., Rimstad, E., Webster, R.G., 2005. Infectious salmon anaemia virus. In Virus Taxonomy. Classification and Nomenclature of Viruses: Eighth Report of the International Committee on Taxonomy of Viruses, pp. 681–693. Edited by C. M. Fauquet, M. A. Mayo, J. Maniloff, U. Desselberger & L. A. Ball. New York: Elsevier, Academic Press.

Kibenge, F.S.B., Lyaku, J.R., Rainnie, D., Hammell, K.L., 2000a. Growth of infectious salmon anaemia virus in CHSE-214 cells and evidence for phenotypic differences between virus strains. J. Gen. Virol. 81, 143-150.

Kibenge, M.J.T., Opazo, B., Rojas, A. Kibenge, F.S.B., 2002. Serological evidence of infectious salmon anaemia virus (ISAV) infection in farmed fishes, using an indirect enzyme-linked immunosorbent assay. Dis. Aquat. Org. 51, 1-11.

Kibenge, F.S.B., Whyte, S.K., Hammell, K.L., Rainnie, D., Kibenge, M.T., Martin, C.K., 2000b. A dual infection of infectious salmon anaemia (ISA) virus and a togavirus-like virus in ISA of Atlantic salmon Salmo salar in New Brunswick, Canada. Dis. Aquat. Organ. 42, 11-15.

Kibenge, F.S., Munir, K., Kibenge, M.J., Joseph, T., Moneke, E., 2004. Infectious salmon anemia virus: causative agent, pathogenesis and immunity. Anim. Health Res. Rev. 5, 65-78.

Kibenge, M.J.T., Opazo, B., Rojas, A. Kibenge, F.S.B., 2002. Serological evidence of infectious salmon anaemia virus (ISAV) infection in farmed fishes, using an indirect enzyme-linked immunosorbent assay. Dis. Aquat. Org. 51, 1–11.

Knottnerus, J.A., Muris, J.W., 2003. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J. Clin. Epidemiol. 56, 1118-1128.

Koren, C.W.R., Nylund, R., 1997. Morphology and morphogenesis of infectious salmon Anaemia virus replicating in the endothelium of Atlantic salmon *Salmo salar*. Dis. Aquat. Org. 29, 99-109.

Lijmer, J.G., Bossuyt, P.M., 2009. Various randomized designs can be used to evaluate medical tests. J. Clin. Epidemiol. 62, 364-373.

Lijmer, J.G., Mol, B.W., Heisterkamp, S., Bonsel, G.J., Prins, M.H., van der Meulen, J.H., Bossuyt, P.M., 1999. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 282, 1061-1066.

Løvdal, T., Enger, O., 2002. Detection of infectious salmon anemia virus in sea water by nested RT-PCR. Dis. Aquat. Organ. 49, 123-128.

Lovely, J.E., Dannevig, B.H., Falk, K., Hutchin, L., MacKinnon, A.M., Melville, K.J., Rimstad, E., Griffiths, S.G., 1999. First identification of infectious salmon anaemia in North America with Haemorrhagic kidney syndrome. Dis. Aquat. Dis. 35, 145-148.

Loy, C.T., Irwig, L., 2004. Accuracy of Diagnostic Tests Read With and Without Clinical Information: A Systematic Review. JAMA 292, 1602-1609.

Maclure, M., Willett, W.C., 1987. Misinterpretation and misuse of the Kappa statistic. Am. J. Epidemiol. 126,161-169.

MacWilliams, C., Johnson, G., Groman, D., Kibenge, F.S., 2007. Morphologic description of infectious salmon anaemia virus (ISAV)-induced lesions in rainbow trout *Oncorhynchus mykiss* compared to Atlantic salmon *Salmo salar*. Dis. Aquat. Organ. 78, 1-12.

Martin, S.W., Shoukri, M., Thorburn, M.A., 1992. Evaluating the health status of herds based on tests applied to individuals. Prev. Vet. Med., 14, 33-43.

Martin, S.W., 2007. The way forward: management of infectious salmon anemia in the bay of fundy. A report prepared for the National Aquatic Animal Health Program of the Canadian Food Inspection Agency. 106 pp.

Maxwell, A.E., 1970. Comparing the classification of subjects by two independent judges. Br. J. Psychiatry 116, 651-655.

McAllister, P.E., Densmore, C.L., Barbash, P.A., 2003. Infectious salmon anaemia virus: injection challenge and waterborne transmission monitored by hematology and polymerase chain reaction assay. In: 28th Annual Eastern Fish HealthWorkshop.

McBeath, A.J.A., Burr, K.L.A., Cunningham, C.O., 2000. Development and use of a DNA probe for confirmation of cDNA from infectious salmon anaemia virus (ISAV) in PCR products. Bull. Eur. Ass. Fish Pathol. 20, 130-134.

McBeath, S.J., Ellis, L.M., Cook, P.F., Wilson, L., Urquhart, K.L., Bricknell, I.R., 2006. Rapid development of polyclonal antisera against infectious salmon anaemia virus and its optimization and application as a diagnostic tool. J. Fish Dis. 29, 293-300.

McCarthy, E.L., Egeler, T.J., Bickerstaff, L.E., Pereira da Cunha, M., Millard, P.J., 2006. Detection and identification of IHN and ISA viruses by isothermal DNA amplification in microcapillary tubes. Anal. Bioanal. Chem. 386, 1975-1984.

McCarthy, E.L., Bickerstaff, L.E., da Cunha, M.P., Millard, P.J. 2007. Nucleic acid sensing by regenerable surface-associated isothermal rolling circle amplification. Biosens. Bioelectron. 22, 1236-1244.

McClure, C.A., Hammell, K.L., Dohoo, I.R., Nerette, P., Hawkins, L.J., 2004. Assessment of infectious salmon anaemia virus prevalence for different groups of farmed Atlantic salmon, Salmo salar L., in New Brunswick. J. Fish. Dis. 27, 375-383.

McClure, C.A., Hammell, K.L., Stryhn, H., Dohoo, I.R., Hawkins, L.J., 2005. Application of surveillance data in evaluation of diagnostic tests for infectious salmon anemia. Dis. Aquat. Org. 63, 119-127.

McKenna, S.L., Dohoo, I.R., 2006. Using and Interpreting diagnostic Tests. Vet. Clin. Food Anim. 22, 195-205.

Melville, K.J., Griffiths, S.G., 1999. Absence of vertical transmission of infectious salmon anemia virus (ISAV) from individually infected Atlantic salmon *Salmo salar*. Dis Aquat Organ. 38, 231-234.

Mikalsen, A.B., Teig, A., Helleman, A.L., Mjaaland, S., Rimstad, E., 2001. Detection of infectious salmon anaemia virus (ISAV) by RT-PCR after cohabitant exposure in Atlantic salmon *Salmo salar*. Dis. Aquat. Org. 47, 175–181.

Miller, W.C., 1998a. Can we do better than discrepant analysis for new diagnostic test evaluation. Clin. Infect. Dis. 27, 1186-93.

Miller, W.C., 1998b. Bias in discrepant analysis: when two wrongs do not make a right. J. Clin. Epidemiol. 51, 219–231.

Mintiens, K., Toft, N., Lewis, F., Verloo, D., Georgiadis, M., Johnson, W., Gardner, I., Wright, P., Gunn, G., Greiner, M., . Guidelines on the use of no-gold standard methods for estimating diagnostic characteristics of microbiological and serological assays. OIE Sci. Tech. Rev. (submitted).

Mjaaland, S., Rimstad, E., Falk, K., Dannevig, B.H., 1997. Genomic characterization of the virus causing infectious salmon anaemia in Atlantic salmon (*Salmo salar* L): an orthomyxo-like virus in a teleost. J. Virol. 71, 7681-7686.

Mjaaland, S., Rimstad, E., Cunningham, C.O., 2002. Molecular diagnosis of infectious salmon anaemia. In: Cunningham, editor. Reviews: Methods and Technology in Fish Biology and Fisheries. London: Kluwer Academic Publishers.

Moneke, E., Groman, D.B., Wright, G.M., Stryhn, H., Johnson, G.R., Ikede, B.O., Kibenge, F.S., 2005. Correlation of virus replication in tissues with histologic lesions in Atlantic salmon experimentally infected with infectious salmon anemia virus. Vet. Pathol. 42, 338-349.

Moneke, E.E., Kibenge, M.J., Groman, D., Johnson, G.R., Ikede, B.O., Kibenge, F.S., 2003. Infectious salmon anemia virus RNA in fish cell cultures and in tissue sections of atlantic salmon experimentally infected with infectious salmon anemia virus. Vet. Diagn. Invest. Sep;15(5):407-17.

Moons, K.G.M., Harrell, F.E., 2003. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad. Radiol. 10, 670-672.

Munir, K., Kibenge, F.S.B., 2004. Detection of infectious salmon anaemia virus by real-time RT-PCR. J. Virological Methods 117, 37–47.

Mullins, J.E., Groman, D., Wadowska, D., 1998. Infectious salmon anaemia in salt water Atlantic salmon (*Salmo salar* L) in New Brunswick, Canada. Bull. Eur. Assoc. Fish. Pathol. 18, 110-114.

Muñoz-Zanzi, C., Thurmond, M., Hietala, S., Johnson, W., 2006. Factors affecting sensitivity and specificity of pooled-sample testing for diagnosis of low prevalence infections. Prev. Vet. Med. 74, 309-322.

Naylor, R. L., Hardy, R.W., Bureau, D.P., Chiu, A., Elliott, M., Farrell, A.P., Forster, I., Gatlin, D.M., Goldburg, R.J., Hua, K., Nichols, P.D., 2009. Feeding aquaculture in an era of finite resources. Proc. Natl. Acad. Sci. 106, 15103-15110.

Neath, A.A., Samaniego, E.J., 1997. On the efficacy of Bayesian inference for nonidentifiable models. Am. Stat. 51: 225-232.

Nérette, P., Dohoo, I., Hammell, L., Gagné, N., 2005a. Estimation of specificity and sensitivity of three diagnostic tests for infectious salmon anaemia virus in the absence of a gold standard. J. Fish Dis. 28, 89-99.

Nérette, P., Dohoo, I., Hammell, L., Gagné, N., Barbash, P., MacLean, S., Yason, C., 2005b. Estimation of the repeatability and reproducibility of three tests for infectious salmon anaemia virus. J. Fish Dis. 28, 101-110.

Nérette, P., Hammell, L., Dohoo, I., Gardner I., 2008a. Evaluation of testing strategies for infectious salmon anaemia and implications for surveillance and control programs. Aquaculture 280, 53-59.

Nérette, P., Stryhn, H., Dohoo, I., Hammell, L., 2008b. Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic tests. Prev. Vet. Med. 85, 207-225.

Nylund, A., Alexandersen, S., Lovik, P., Jakobsen, P., 1994. Mechanism fro transmission of infectious salmon anaemia (ISA). Dis. Aquat. Organ. 19, 95-100.

Nylund, A., Hovland, T., Watanabe, K., Endersen, C., 1995. Presence of infectious salmon anaemia virus (ISAV) in tissues of Atlantic salmon, Salmo salar L., collected during three separate outbreaks of the disease. J. Fish Dis. 18, 135-145.

Obuchowski, N.A., 1998. Sample Size Calculations in Studies of Test Accuracy. Statist. Meth. Medic. Res. 7, 371-392.

Office International des Epizooties Website: http://www.oie.int

Office International des Epizooties, 2008. OIE Quality Standard and Guidelines for Veterinary Laboratories: Infectious Diseases. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 70pp.

Office International des Epizooties, 2009a. OIE Aquatic Animal Health Code. 12th Edition. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 99-104.

Office International des Epizooties, 2009b. Manual of Diagnostic Tests for Aquatic Animals 2009. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 10-30.

Olsen, Y.A., Falk, K., Reite, O.B., 1992. Cortisol and lactate levels in Atlantic salmon *Salmo salar* developing infectious anaemia (ISA). Dis. Aquat. Organ. 14, 99-104.

Opitz, H.M., Bouchard, D., Anderson, E., Blake, S., Nicholson, B., Keleher, W., 2000. A comparison of methods for the detection of experimentally induced subclinical infectious salmon anaemia in Atlantic salmon. Bull. Eur. Assoc. Fish. Pathol. 20, 12-22.

Pearson, K., 1900. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London, Series A , vol. 195, 1-47.

Jane F. Pendergast, Stephen J. Gange, Mary J. Lindstrom, 2005. Correlated Binary Data. Standard Article (http://www.mrw.interscience.wiley.com/emrw/9780470011812/eob/article/b2a10018/current/pdf) , Encyclopedia of Biostatistics, John Wiley & Sons, Inc. 16 pp.

Plarre, H., Devold, M., Snow, M., Nylund, A., 2005. Prevalence of infectious salmon anaemia virus (ISAV) in wild salmonids in western Norway. Dis. Aquat. Organ. 66, 71-79.

Pouillot, R., Gerbier, G., Gardner, I.A., 2002. "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. Prev. Vet. Med. 53, 67-81.

Ransohoff, D.F., Feinstein, A.R., 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N. Engl. J. Med. 17, 926-930.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009) *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Reitsma, J.B., Rutjes, A.W.S., Khan, K.S., Coomarasamy, A., Bossuyt, P.M., 2009. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J. Clin. Epidemiol. 62, 797-806.

Rimstad, E., Falk, K., Mikalsen, A.B., Teig, A., 1999. Time course tissue distribution of infectious salmon anaemia virus in experimentally infected Atlantic salmon Salmo salar. Dis. Aquat. Organ. 36: 107–112.

Rimstad, E., Mjaaland, S., Snow, M., Mikalsen, A.B., Cunningham, C.O., 2001. Characterization of the infectious salmon anaemia virus genomic segment that encodes the putative hemagglutinin. J. Virol. 75, 5352-5356.

Rimstad, E., Mjaaland, S., 2002. Infectious salmon anemia virus: an orthomyxovirus causing an emerging infection in Atlantic salmon. APMIS 110, 273-282.

Rodger, H.D., Turnbull, T., Muir, F., Millar, S., Richards, R., 1998. Infectious salmon anaemia (ISA) in United Kingdom. Bull. Eur. Assoc. Fish. Pathol. 18, 115-116.

Rogan, W.J., Gladen, B., 1978. Estimating prevalence from the results of a screening test. Am. J. Epidemiol. 107, 71-76.

Rolland, J.B., Bouchard, D., Coll, J., Winton, J.R., 2005. Combined use of ASK and SHK-1 cell lines to enhance the detection of infectious salmon anemia virus. J. Vet. Diagn. Invest. 17, 151-157.

Rolland, J.B., Bouchard, D.A., Winton, J.R., 2003. Improved diagnosis of infectious salmon anaemia virus by use of a new cell line derived from Atlantic salmon kidney tissue. In: Miller O and Cipriano RC, technical coordinators. International Response to Infectious Salmon Anemia: Prevention, Control, and Eradication: Proceedings of a Symposium, 3–4 September 2002, New Orleans, LA. Technical Bulletin 1902. Washington, DC: Department of Agriculture, Animal and Plant Health Inspection Service, pp. 63–68.

RSE (the Royal Society of Edinburgh), 2002. The scientific issues surrounding the control of infectious salmon anaemia (ISA) in Scotland. A Report of the Royal Society of Edinburgh Working Party on Infectious Salmon Anaemia. http://www.rse.org.uk/enquiries/isa/report.pdf. 25pp.

Salonius, K., Barton, T., Erdal, J.I., 2003. Efficacy of a multivalent vaccine against cohabitation challenge with Norwegian ISAV. In: 3rd International Symposium on Fish Vaccinology, Bergen, Norway, April 9-11, 2003.

Sanchez, L., Abuin, M., Amaro, R., 1993. Cytogenetic characterization of the AS cell line derived from the Atlantic salmon (*Salmo salar* L.). Cytogen. Cell Genet. 64, 35-38.

Schulzer, M., Anderson, D.R., Drance, S.M,. 1991. Sensitivity and specificity of a diagnostic test determined by repeated observations in the absence of an external standard. J. Clin. Epidemiol. 44, 1167-1179.

Shapiro, D.E., 1999. The interpretation of diagnostic tests. Stat. Methods Med. Res. 8, 113-134.

Simel, D.L., Samsa, G.P., Matchar, D.B., 1991. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J. Clin. Epidemiol. 44, 763-770.

Simko, E., Brown, L.L., MacKinnon, A.M., Byrne, P., Ostland, V.E., Ferguson, H.W., 2000. Experimental infection of Atlantic salmon (*Salmo salar* L.) with infectious salmon anaemia virus - histopathological study. J. Fish Dis. 23, 27-32.

Simko, E., Falk, K., Poppe, T.T., Ferguson, H.W., 2001. Plasma protein changes in Atlantic salmon, Salmo salar L., with infectious salmon anaemia. J. Fish Dis. 24, 293-298.

Smail, D., Grant, R., Ross, K., Bricknell, I.R., Hastings, T.S., 2000. The use of haemadsorption for the isolation of infectious salmon anaemia virus on SHK-1 cells from Atlantic salmon (*Salmo salar* L.) in Scotland. Bull. Eur. Ass. Fish Pathol. 20, 212-214.

Snow, M., McKay, P., Matejusova, I., 2009. Development of a widely applicable positive control strategy to support detection of infectious salmon anaemia virus (ISAV) using Taqman real-time PCR. J. Fish Dis. 32, 151-156.

Snow, M., McKay, P., McBeath, A.J.A., Black, J., Doig, F., Kerr, R., Cunningham, C.O., Nylund, A., Devold, M., 2006. Development, application and validation of a Taqman real-time RT-PCR assay for the detection of infectious salmon anaemia virus (ISAV) in Atlantic salmon (*Salmo salar*). Devel. Biolog. 126, 133-145.

Snow, M., Raynard, R.S., Murray, A.G., Bruno, D.W., King, J.A., Grant, R., Bricknell, I.R., Bain, N., Gregory, A., 2003. An evaluation of current diagnostic tests for the detection of infectious salmon anaemia virus (ISAV) following experimental water-borne infection of Atlantic salmon, Salmo salar L. J. Fish. Dis. 26, 135-45.

Sommer, A.-I., Mennen, S., 1996. Propagation of infectious salmon anaemia virus in Atlantic salmon (*Salmo salar* L.) head kidney macrophages. J. Fish Dis. 19, 179-183.

Sommer, A.I., Mennen, S., 1997. Multiplication and haemadsorbing activity of infectious salmon anaemia virus in the established Atlantic salmon cell line. J. Gen. Virol. 78, 1891-1895.

Speilberg , L., Evensen, O., Dannevig, B.H., 1995. A sequential study of the light and electron microscopical liver lesions of infectious anemia in Atlantic salmon (*Salmo salar* L). Vet. Pathol. 32, 466-478.

Spiegelhalter, D.J., Thomas, A., Best, N.G., 2003. WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit.

Stagg, R.M., Bruno, D.W., Cunningham, C.O., Raynard, R.S., and 6others (2001) Epizootiological investigations into an outbreakof infectious salmon anaemia (ISA) in Scotland. FRSMarine Laboratory Report No 13/01, Fisheries ResearchServices, Marine Laboratory, Aberdeen.

Starkey, W.G., Smail, D.A., Bleie, H., Muir, K.F., Ireland, J.H., Richards, R.H., 2006. Detection of infectious salmon anaemia virus by real-time nucleic acid sequence based amplification. Dis. Aquat. Organ. 72, 107-113.

Stuart, A., 1955. A test for homogeneity of the marginal distribution in a two-way classification. Biometrika 42, 412-416.

Thorburn, M.A., 1999. Applying epidemiology to infectious diseases of fish. 689-722. in P. T. K. Woo and D. W. Bruno, editors.

Thorud, K.E., Djupvik, H.O., 1988. Infectious salmon anaemia in Atlantic salmon (*Salmo salar* L). Bull. Eur. Assoc. Fish. Pathol. 8, 109-111.

Thorud, K., 1991. Infectious salmon anaemia. Transmission trials. Haematological, clinical chemical and morphological investigations. Dr. Scient. thesis, Norwegian College of veterinary medicine.

Toft, N., Jorgensen, E., Hojsgaard, S., 2005. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev. Vet. Med. 68, 19–33.

Toft, N., Innocent, G., Gettingby, G., Reid, S., 2007. Assessing the convergence of Markov Chain Monte Carlo methods: An example from evaluation of diagnostic tests in absence of a gold standard. Prev. Vet. Med. 79, 244-256.Agresti, A., 2002. Categorical data analysis. New York: Wiley.

Torrance-Rynard, V.L., Walter, S.D., 1997. Effects of dependent errors in the assessment of diagnostic test performance. Stat. Med. 16, 2157-2175.

Totland, G.K., HjeItnes, B.K., Flood, P.R., 1996. Transmission of infectious salmon anaemia (ISA) through natural secretions and excretions from infected smelts of Atlantic salmon *Salmo salar* during their presymptomatic phase. Dis. Aquat. Org. 26, 25-31.

Uebersax, J., 2007. Statistical methods for rater agreement. Website: www.john-uebersax.com/stat/agree.html .

Van den Bruel, A., Aertgeerts, B., Buntinx, F. 2006. Results of diagnostic accuracy studies are not always validated. J. Clin. Epidemiol. 59, 559-566.

Van den Bruel, A., Cleemput, I., Aertgeerts, B., Ramaekers, D., Buntinx, F., 2007. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J. Clin. Epidemiol. 60,1116-1122.

Van der Veen, A.M.H., Pauwels, J., 2000. Uncertainty calculations in the certification of reference materials, 1: principles of analysis of variance. Accredit. Qual. Assur. 5 , 464–469.

Van der Veen, A.M.H., Linsinger, P.J.T., Pauwels, J., 2001a. Uncertainty calculations in the certification of reference materials, 2: homogeneity study. Accredit. Qual. Assur. 6, 26–30.

Van der Veen, A.M.H., Linsinger, P.J.T., Lamberty, A., Pauwels, J., 2001b. Uncertainty calculations in the certification of reference materials, 3: stability study. Accredit. Qual. Assur. 6, 257–263.

Vecchio, T.J., 1966. Predictive value of a single diagnostic test in unselected populations. N. Engl. J. Med. 274, 1171-1173.

WAHID, World Animal Health Information Database Interface for OIE. Website: http://www.oie.int/wahis/public.php?page=disease_status_map&disease_type=Aquatic&disease_id=160&sta_method=semesterly&selected_start_year=2008&selected_report_period=1&selected_start_month=1&page=disease_ status_map.

Wergeland, H.I., Jakobsen, R.A., 2001. A salmonid cell line (TO) for production of infectious salmon anaemia virus (ISAV). Dis. Aquat. Organ. 44, 183-190.

Wilson, I.G., 1997. Inhibition and facilitation of nucleic acid amplification. Appl. Environ. Microbiol. 63, 3741-3751.

Wilson, L., McBeath, S.J., Adamson, K.L., Cook, P.F., Ellis, L.M., Bricknell, I.R., 2002. An alkaline phosphatase-based method for the detection of infectious salmon anaemia virus (ISAV) in tissue culture and tissue imprints. J. Fish Dis. 25, 615-619.

Wong, M.L., Medrano, J.F., 2005. Real-time PCR for mRNA quantitation. Biotechniques 39, 75-85.

Workenhe, S.T., Wadowska, D.W., Wright, G.M., Kibenge, M.J., Kibenge, F.S., 2007. Demonstration of infectious salmon anaemia virus (ISAV) endocytosis in erythrocytes of Atlantic salmon. Virol J. 25, 4-13.

Workenhe, S.T., Kibenge, M.J., Iwamoto, T., Kibenge, F.S., 2008. Absolute quantitation of infectious salmon anaemia virus using different real-time reverse transcription PCR chemistries. J. Virol. Methods. 154, 128-134.

Wright, P., Edwards, S., Diallo, A., Jacobson, R., 2006. Development of a framework for international certification by OIE of diagnostic tests validated as fit for purpose. Dev. Biolog. 126, 43-51.

Yerushalmy, J., 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Publ. Health Rep. 62, 1432-1449.

Youden, M.H., 1950. Index for rating diagnostic tests. Cancer 3, 32-35.

Zhou, X.H., 1993. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Com. Statist. Theor. Meth. 22, 3177-3198.

Zhou, X.H., 1994. Effect of verification bias on positive and negative predictive values. Statist. Med. 13, 1737-1745.

Zhou, X.H., 1998. Correcting for verification bias in studies of a diagnostic test's accuracy. Stat. Methods Med. Res. 7, 337-353.

Chapter II: TRADITIONAL DESCRIPTIVE ANALYSIS AND NOVEL VISUAL REPRESENTATION OF DIAGNOSTIC REPEATABILITY AND REPRODUCIBILITY: APPLICATION TO AN INFECTIOUS SALMON ANAEMIA VIRUS RT-PCR ASSAY

Abstract

As a component of diagnostic test evaluation, the estimation of repeatability and reproducibility of an assay is necessary to assess the robustness and the transferability of the method among laboratories. Respectively defined as the agreement within and between laboratories, repeatability and reproducibility of a qualitative diagnostic test are traditionally reported using observed proportion of agreement or Kappa values. Applied to a recently designed RT-PCR assay for the detection of infectious salmon anaemia virus, repeatability within a national reference laboratory and reproducibility with two independent regional laboratories were investigated. Additionally, homogenization of fish kidney tissue was conducted to potentially provide more uniform submission material, and to assess the effect of homogenization on laboratory comparability. Comparison of agreement between non-homogenized and homogenized tissue samples revealed different patterns of test results and unexpected alterations of agreement due to homogenization. This observation may be explained by cross-contamination of some samples during the homogenization process. One of the laboratories was in clear disagreement with the two others and impacted the overall reproducibility of the assay. Agreement levels were visually described using a novel tree-shape representation inspired from phylogenetic studies. The resulting phylogram illustrated the proximity of test findings between repeated samples within a laboratory and between laboratories, and facilitated the interpretation of the agreement levels.

*This chapter was published in **Preventive Veterinary Medicine**: 2009; 92(1-2):9-19.*

2.1. Introduction

2.1.1 Traditional evaluation of diagnostic precision

The Office International des Epizooties (OIE, World Organisation for Animal Health) aims to safeguard international trade by publishing standards and guidelines for health and self-declaration of disease-freedom for animals and animal products. To diagnose infectious diseases and associated pathogens, the OIE recommends use of certified or validated diagnostic assays (OIE, 2008). Diagnostic validation is defined as the evaluation of a test method based on its fitness for a specific purpose (OIE, 2008). Validation is a multiple-stage process that determines the operating characteristics of the test including the assessment of its characteristics and performance at the bench level (estimation of analytical sensitivity, specificity, and repeatability), evaluation of its accuracy (diagnostic sensitivity and specificity), and estimation of its precision ("diagnostic" repeatability and reproducibility) at the population level.

The estimation of the test precision is an important step of the validation process, although sometimes overlooked and neglected. Diagnostic repeatability is defined as the variation in test results that are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time (within-laboratory consistency). Diagnostic reproducibility is defined as the variation in test results that are obtained with the same method on identical test items in different laboratories with different operators using different equipment (between-laboratory consistency) (IS0 5725-1, 1994).

The concept of variation in binary outcome diagnostics is associated with the concept of agreement between test runs. We defined as "test run" or "run" a set of results obtained using the same method under defined conditions relative to the testing laboratory and the nature of the sample (identical conditions for repeatability and similar conditions for reproducibility). Agreement is traditionally expressed using the proportion of agreement (Pa) (proportion of tests results that agree) or using Cohen's Kappa values (κ) (Dohoo et al., 2009). It has been suggested that precision for binary tests can also be assessed using predictive intervals of diagnostic sensitivity (DSe), specificity (DSp) or overall accuracy (Cleophas et al., 2008). This study was restricted to the evaluation of diagnostic repeatability and reproducibility to suit the requirements of international standards (OIE, 2008).

2.1.2 Novel approach to diagnostic precision, inspired by phylogenetics

Phylogenetics is a discipline that investigates the relationship among organisms according to their gene similarity. The pairwise comparison of aligned nucleotide or amino acid sequences determines the degree of similarity (agreement) between genes. The measure of similarity is calculated as the proportion of nucleotides, or amino acids, that are identical between two sequences (Vandamme, 2003). Proportions of similarity (or dissimilarity) are usually summarized in a pairwise genetic distance matrix. The distance matrix is then used to reconstruct a phylogenetic tree that illustrates the evolutionary relationship among compared organisms. Distance matrices are comparable to agreement matrices that are reported in diagnostic evaluation studies. Methods using

distance matrices for phylogenetic tree inferences are referred to as distance-based methods in contrast with character-based methods that integrate additional character information. Genes with high sequence agreement will be positioned closer to each other, whereas genes with low sequence agreement will not group together. Similarly to genetic sequences, laboratory test results can be aligned and analyzed using distance-matrix based models to visually represent agreement among laboratories in a tree shape.

2.1.3 Infectious salmon anaemia virus

Infectious Salmon Anaemia virus (ISAV) is an Orthomyxovirus, genus *Isavirus,* causing a hemorrhagic syndrome in salmonids. Primarily pathogenic for Atlantic salmon, *Salmo salar* L., the viral agent, causing high mortality, is a serious threat to the economic sustainability of many salmon aquaculture industries worldwide. Originally observed in Norway in 1984 (Thorud & Djupvik, 1988), clinical ISA was then chronologically reported in Canada (Mullins et al., 1998), Scotland (Rodger et al., 1998), Faroe Islands (Anonymous, 2000), USA (Bouchard et al., 2001), and recently in Chile (Godoy et al., 2008). Absent in some areas of Atlantic salmon production (e.g. Tasmania, Australia; British Columbia, Canada), ISAV is listed as a reportable aquatic disease by the OIE (OIE, 2009). Consequently, for international trade purposes, diagnostic methods used for screening, certification, confirmation, and control require validation. The implementation of the National Aquatic Animal Health Program, including national reference laboratories and surveillance programs, aims at controlling and preventing the emergence and spread of aquatic disease in Canada. Since ISAV surveillance is a goal of the program, it was

required that a recently designed Reverse-Transcriptase Polymerase Chain Reaction (RT-PCR) assay for ISAV (Gagné et al., data unpublished) be validated.

2.1.4 Repeatability and reproducibility of the ISAV RT-PCR

A single study previously investigated the repeatability and reproducibility of an ISAV RT-PCR assay in three different laboratories (Nérette et al., 2005). The study revealed substantial differences in repeatability (Pa ranging from 76 to 98% and κ from 0.50 to 0.96). In addition, there was a serious disagreement explained by one laboratory with a higher proportion of positive tests although a substantial reproducibility was found between the two others (Pa = 91% and κ = 0.79). The authors proposed that factors associated with sample and testing conditions may have affected the assessment of reproducibility and repeatability: i) heterogeneous distribution of virus in the organ may have resulted in virus quantity inconsistency among replicated samples; ii) differences in testing protocols (i.e. different sets of primers and methods) may have compromised the comparability of laboratories; and iii) differences in agarose gel interpretation and confirmation protocols (i.e. whether a sample with a weak gel band was retested) may have also affected interpretation of results differently in the three laboratories.

2.1.5 Objectives

The objective of this study was three-fold. The first objective was to describe qualitative diagnostic precision of a newly designed ISAV RT-PCR (Gagné et al., data

unpublished) in three different laboratories using identical standard operating procedures for testing and interpretation. Specifically, we estimated the repeatability only within the designated National Reference Laboratory for ISAV in Canada (the Aquatic Animal Health Section of the Gulf Fisheries Centre, Department of Fisheries and Oceans, Moncton, Canada), and the reproducibility by including two other independent laboratories in the study. The second objective was to investigate the impact of the potential heterogeneous distribution of viral particles among replicate samples by assessing agreement of homogenized tissue samples. The third objective was to develop a novel visual approach to describe test agreement using distance-matrix based tree reconstruction inspired from phylogenetic studies. This new approach was not intended to replace former methods but to facilitate the illustration and complement interpretation of agreement with a new perspective.

2.2. Material and methods

2.2.1 Study material

2.2.1.1 Sample selection

Kidney samples from 100 Atlantic salmon were selected from archives of different sources to target a prevalence of approximately 50% according to McClure et al. (2004); briefly, 45 apparently healthy fish (15 fish from each cage) were from exposed cages in 3 different infected sites (expected prevalence of 28.1%), 35 apparently healthy fish were from a known infected cage (expected prevalence of 41.5%), and 20 dead or

moribund fish (10 fish from two different sites) were from ISA clinically affected cages (expected prevalence of 100%). Kidney samples were collected aseptically from each fish in replicates of six and stored in RNA*later* (Ambion Inc., Austin, TX, USA) at -80 °C after a 24 hour period at 4 °C.

2.2.1.2 Sample allocation

Each sample was coded with a random identification number to blind laboratory operators and to avoid test-review bias (Ransohoff and Feinstein, 1978). Sample distribution and testing objectives are summarized in Fig. 2.1. From each salmon, duplicate samples (t1, t2) were sent on dry ice to the reference laboratory (lab A) to estimate repeatability, and single samples (t) were transported on dry ice to two other laboratories (labs B and C) to estimate reproducibility. The remaining two samples were combined, homogenized and aliquoted with equal volume (250 µL) in four coded microtubes.

Homogenization was performed in lab A by transferring the two samples (h1, h2) in a 2 ml microtube filled to the upper limit with RNA*later* and homogenized using a FastPrep® FP120A homogenizer (MP Biomedicals) at 5.5 m/sec for 20 s twice. Two aliquots (h) of 250 µl of homogenate from each fish stayed in lab A stored at -80 °C to estimate the repeatability, and single aliquots of 250 µl of homogenates were sent frozen on dry ice to the other two laboratories to estimate the reproducibility (Fig. 2.1). Each of the participating laboratories agreed to test for the presence and absence of ISAV using the same RT-PCR protocol provided by the reference laboratory (lab A).

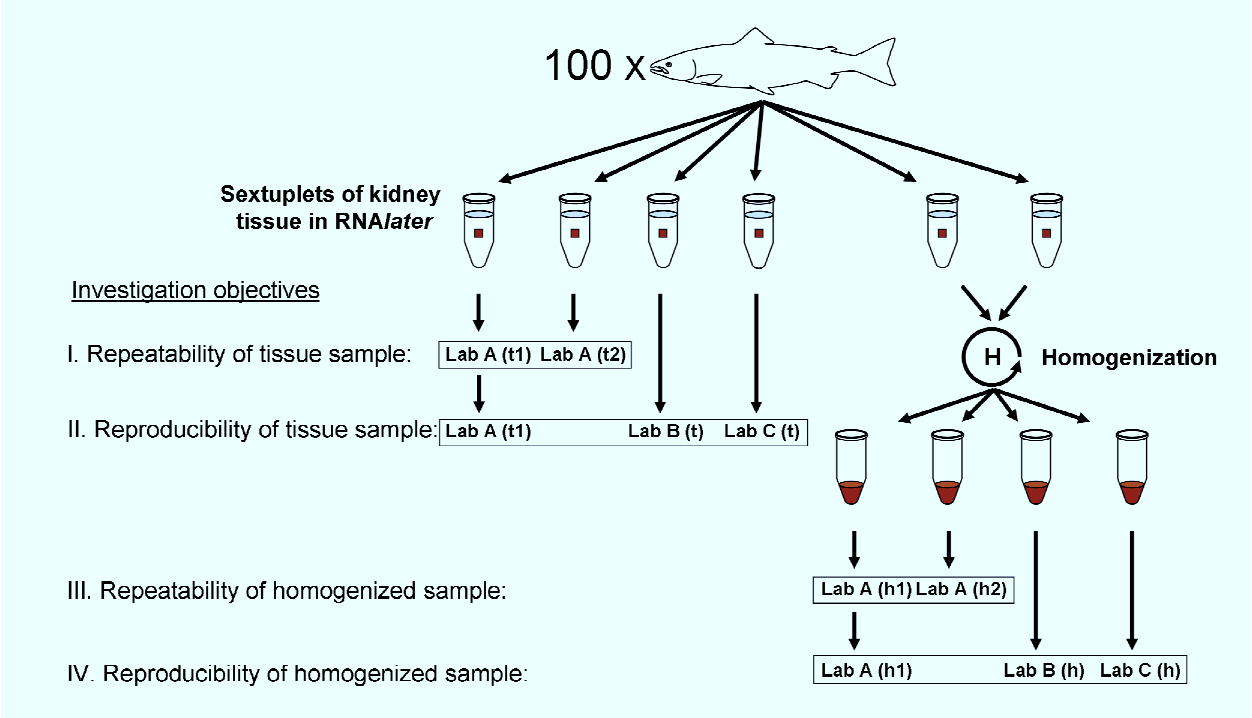


Fig. 2.1. Sample allocation and investigation objectives to study RT-PCR repeatability and reproducibility. (t1): non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample.

2.2.2 Testing protocol (Reverse-Transcriptase Polymerase Chain Reaction and interpretation)

For RNA extraction, a piece of tissue (approximately 30 ± 5 mg) was removed and homogenized in 1 ml of TRI reagent (Molecular Research Inc) with a FastPrep FP120 (Savant Instruments). For homogenates, microtubes were centrifuged to remove the RNA*later* before adding 1 ml of TRI reagent and homogenization. Manufacturer's instructions were followed, except for 2 additional washes of the RNA pellet with 75% ethanol. RNA pellets were resuspended in 50 µl of sodium citrate buffer 1mM pH 6.4 containing an RNase inhibitor (Qiagen). RNA was further diluted if necessary, and quantified on a spectrophotometer and normalized. A maximum of 1000 ng/µl was used for reverse transcription.

A one-step RT-PCR was used to detect ISAV, using the Qiagen One-step RT-PCR kit (Qiagen). The mixture comprised 5 µl of Q solution, 0.32 µM of each primer (404F : 5' tgg gca atg gtg tat ggt atg a-3' and RA3(583R): 5' gaa gtc gat gaa ctg cag cga-3'), 1 µl of enzyme, 5 µl of buffer, 1 µl of dNTP, 11.2 µl of H_2O and ≤ 1 µg of RNA, for a total volume of 25 µl. PCR conditions consisted of an initial hold at 50 °C for 30 min, and 95 °C for 15 min, followed by 10 cycles of touchdown PCR starting with 94 °C for 40 s, 72 °C for 40 s, and 72 °C for 60 s, and lowering by 1 °C the annealing temperature after each cycle. Then 40 cycles at 94 °C for 40 s, 62 °C for 40 s, and 72 °C for 60 s were added, and a final extension of 72 °C for 10 min, and holding at 20 °C completed the program. PCR products (10 µl) were electrophoresed in 6% acrylamide, visualized with

ethidium bromide, and compared to positive controls and a DNA ladder. A band at the same expected size (179 bp) as the positive control was considered positive. For quality control, extraction blanks (no sample) were included every 15th tube during extractions, and blanks (water) were added at the RT-PCR step. Electrophoresis gels were examined carefully, and PCR was repeated on samples where a very weak intensity band at the expected size was observed initially. If the second PCR result was positive again, the final result was positive; if not, it became negative.

2.2.3 Statistical Analysis

2.2.3.1 Descriptive Statistics

Test results from each laboratory were collated and first analyzed using Stata SE 10.0 (Stata Corp., College Station, TX, USA, 2007). The first agreement statistic computed was the observed proportion of agreement (Pa), giving the proportion of paired test results that agreed either on a positive or on a negative test result between two test runs. Exact confidence intervals (CIs) for observed agreement were computed. An average Pa was also computed as a mean of Pa estimates from all possible pairs of test runs within lab A (overall repeatability), and among the three laboratories (average reproducibility of homogenized and non-homogenized samples, and overall reproducibility). CIs were computed using 5% and 95% percentile values from bootstrapped estimates resampled 1,000 times.

The second agreement statistic computed was Cohen's Kappa (κ), commonly used for subjective rating. Ranging from -1 to +1, this value represents the level of

agreement beyond agreement expected due to chance (Dohoo et al., 2009). The κ was also computed for agreement among three laboratories data together (Fleiss et al., 2003). CIs for the κ statistic were computed using an analytical method for comparison of two test runs (Fleiss et al., 2003) and a bootstrap method for three runs (Lee and Fung, 1993). Prior to each paired κ estimation, a McNemar's test (exact binomial test for correlated proportions) was performed to assess if proportions of positive results differed between test runs. Evidence of proportion disagreement between runs would constitute disagreement between runs and reduce the interest in κ estimation (Dohoo et al., 2009).

Due to violation of the assumption of independent observations (test results obtained from the same fish), two Pa from different conditions (i.e. repeatability of non-homogenized vs. homogenized samples) could be compared using a McNemar's test by defining agreement/non-agreement (e.g. comparing two runs) as a binary outcome. However, this method does not consider proportion of agreement on a positive and on negative result. The contingency table can contain more than two categories (e.g. agree on a positive result, disagree and agree on a negative result). The effect of homogenization on agreement was then assessed by testing symmetry and marginal homogeneity in contingency tables. The symmetry test compares symmetrical cells around the agreement diagonal of the contingency table, whereas the marginal homogeneity test compares the marginal distributions of the positive test results (Table 2.1). Agreement among all test results from non-homogenized samples was compared to agreement among all test results from homogenized samples using the exact test for symmetry and Stuart-Maxwell test for marginal homogeneity (`symmetry` command; Stata Base Reference Manual, 2007). In addition, following the approach outlined in

Agresti (2002), a quasi-symmetry model for the contingency table was fit, and marginal homogeneity was tested using a likelihood ratio test of symmetry in this model. The same approach was used to compare agreement in homogenized and non-homogenized samples within- and between-laboratory data.

2.2.3.2 Distance matrix

A summary matrix of observed agreement (Pa) and disagreement (i.e. proportion of results that disagree between the two test runs: 1 - Pa) was generated for all possible pairwise comparisons. In phylogenetic methods, the observed disagreement is also called *observed distance* or *p-distance* (Van de Peer and Salemi, 2003). Thus the distance matrix summarized the relative distance of the runs to each other based on their test results. Smaller distance values indicate closer result findings between two laboratories.

2.2.4 Test run phylogram

2.2.4.1 Pseudogold standard

A pseudogold standard (PGS) was created to provide a consensus reference baseline for the test results alignment. Adapting the PGS definition of Nérette et al. (2008), ISA positive and ISA negative classification criteria were arbitrarily based on the combination of six test results for each fish, excluding the duplicate sample results in lab A (3 laboratories testing a non- and a homogenized samples each). "Infected" fish were any

Table 2.1

Contingency table comparing non-homogenized and homogenized sample results from the four tests (2 tests in reference lab A and one test each in participating lab B & C)[a,b]. The number (#) of positive test result "0" and "4" correspond to complete agreement among all samples and laboratories on a negative and on a positive test result respectively, while intermediate categories correspond to the number of positive test result among the four sub-samples regardless of the sample and the laboratory.

		Homogenized					
	# of Positive	0	1	2	3	4	*Marginal*
Non-homogenized	0	**23**	6	6	0	0	*35*
	1	4	**6**	4	1	3	*18*
	2	1	0	**2**	1	2	*6*
	3	0	0	0	**3**	4	*7*
	4	0	0	0	2	**31**	*33*
	Marginal	*28*	*12*	*12*	*7*	*40*	99

[a] **Symmetry test** compared symmetrical cells around the agreement diagonal (in bold)

[b] **Marginal homogeneity test** compared the marginal distributions of non-homogenized and homogenized samples (italicized)

fish with more than three positive tests out of the six (> 3/6). “Non-infected” fish were any fish with three or less positive tests out of the six (≤ 3/6).

2.2.4.2 Alignment formatting

Initially formatted with individuals in rows and runs in columns, tests results were transposed so individuals were in columns and test runs in rows. Negative results, “*0*”, were recoded with an “*a*” (corresponding to adenine) and positive results, “*1*”, were recoded with a “*g*” (corresponding to guanine) in a FASTA format to suit the DNA sequence alignment editor software BioEdit version 7.07 (Hall, 1999) requirements. Test results were edited and displayed as a sequence alignment where only results in disagreement with the PGS are highlighted, and test results in agreement were symbolized by a “.” as a placeholder.

2.2.4.3 Distance-matrix based tree reconstruction model

To conduct tree reconstruction, the FASTA alignment was transferred into the MEGA format using the package MEGA version 4 (Tamura et al., 2007). The alignment was considered as a non protein-coding nucleotide sequence and phylograms were obtained using the distance-based Neighbor-Joining (NJ) method. The model used distances based on the number of differences, and missing data were handled by pairwise deletion. Statistical support for tree topologies were bootstrap-resampled 1,000 times (Felsenstein, 1985). Bootstrap support values (proportion of resampled trees that include the node of interest) were reported in percentage on the nodes of the original tree.

Phylograms were edited using the TreeExplorer software appended to the MEGA package.

2.3. Results

2.3.1 Descriptive statistics

Results were obtained for all eight test conditions with all 100 samples, except that lab C had insufficient material for one homogenized sample (only 99 results for homogenates). Among the 100 non-homogenized samples, duplicates t1 and t2 of lab A detected respectively 42 and 44 positives, and lab B detected 43, while lab C detected 58. Among the 100 homogenized samples, duplicates h1 and h2 of lab A detected respectively 53 and 57 positives, and lab B detected 48, while lab C detected 62 (out of 99 results). Agreement statistics of interest (Pa and κ) and CIs are summarized in Table 2.2.

Overall repeatability revealed slightly lower Pa than overall reproducibility (0.81 and 0.82, respectively), although the overlapping of CIs provided little evidence of significant difference (Table 2.2). Tests from pairwise comparisons involving lab C showed serious disagreement with the two other laboratories regardless of the sample type (significant McNemar's test). Estimates of κ ranged from 0.57 to 0.73 and concurred with Pa results (Table 2.2).

The average proportion of positive results for non-homogenized samples was 46.8%; and the average proportion of positive results for homogenized samples was

Table 2.2

Summary of ISAV diagnostic test descriptive agreement statistics, proportions, and Kappa values, according to sample type and laboratories comparison. t1: reference laboratory A, tissue sample, duplicate 1; t2: reference laboratory A, tissue sample, duplicate 2; h1: reference laboratory A, homogenized sample, duplicate 1; h2: reference laboratory A, homogenized sample, duplicate 2; tB: laboratory B, tissue sample; tC: laboratory C, tissue sample; hB: laboratory B, homogeized sample; hC: laboratory C, homogenized sample.

Agreement level	**Repeatability**		**Reproducibility**					
Sample type	**Non-homogenized**	**Homogenate**	**Non-homogenized**			**Homogenate**		
Lab Comparison	**t1 / t2**	**h1 / h2**	**t1 / tB**	**t1 / tC**	**tB / tC**	**h1 / hB**	**h1 / hC**	**hB / hC**
0 - 0	49	35	51	39	40	40	31	35
1 - 1	35	45	36	39	41	41	46	45
1 - 0	7	8	6	3	2	12	6	2
0 - 1	9	12	7	19	17	7	16	17
Total (count)	100	100	100	100	100	100	99	99
Pa (CI)	**0.84** (0.75-0.90)	**0.80** (0.71-0.87)	**0.87** (0.79-0.93)	**0.78** (0.69-0.86)	**0.81** (0.72-0.88)	**0.81** (0.72-0.88)	**0.78** (0.68-0.85)	**0.81** (0.72-0.88)
Pa average among runs* (CI)	**0.81** (0.75-0.86)		**0.82** (0.76-0.88)			**0.82** (0.77-0.86)	**0.80** (0.74-0.86)	
McNemar's Test (*P*-value)	0.610	0.370	0.780	0.000**	0.000**	0.250	0.033**	0.000**
Kappa (Cohen's)	**0.67**	**0.60**	**0.73**	**0.57**	**0.62**	**0.62**	**0.55**	**0.63**
CI	0.53 - 0.82	0.44 - 0.75	0.60 - 0.87	0.42 - 0.72	0.47 - 0.77	0.47 - 0.77	0.40 - 0.71	0.48 - 0.77
3-replicate Kappa	**na**	**na**	**0.64**				**0.60**	
CI (Bootstrap=1000)	na	na	0.52-0.76				0.47-0.71	

* Computed as the mean of all possible Pa estimates between runs within lab A or among the 3 laboratories

** Significant McNemar's test ($P < 0.05$): significant difference of proportion of positive results between the two test runs; thus corresponding Kappa value is less revelant

na: not applicable

Pa: observed proportion of agreement

CI: confidence interval

56.4%. Table 2.1 shows the contingency table of number of positive test results comparing non-homogenized and homogenized samples. Both symmetry and marginal homogeneity tests showed a significant difference ($P < 0.05$) in overall agreements, repeatabilities and reproducibilities between non-homogenized and homogenized sample results. As an example, for overall agreement, we observed more complete agreements (all four tests agree for a given sample type) on positive results (n =40) than on negative results (n= 28) in homogenized samples whereas non-homogenized samples showed the opposite pattern (33 vs. 35) (Table 2.1). Intermediate marginal proportions (two to three tests that agree) were, however, quite comparable (Table 2.1). Quasi-symmetry modelling procedure showed a good fit to the data and the hypothesis of marginal homogeneity was significantly rejected against that model for overall agreement, repeatability and reproducibility data (all $P < 0.05$).

Table 2.3 represents a summary matrix of observed agreement (Pa) and disagreement (1 - Pa) of all possible pairwise comparisons. The minimum disagreement or distance (0.09) was observed between lab A, duplicate 2, and lab B with non-homogenized sample; and the maximum was observed between non-homogenized sample in lab A and homogenized sample in lab C (0.25). Additionally, significant McNemar's test revealed serious disagreement despite high Pa for several pairwise comparisons (Table 2.3).

2.3.2 Test runs phylogram

According to the PGS, out of the 100 salmon sampled, 48 were positive and 52

Table 2.3

Agreement matrix with proportion of agreement (lower left corner) **and proportion of disagreement or distance** (top right corner in bold) **between runs**. (t1): non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample

Runs	LabA(t1)	LabA(t2)	LabB(t)	LabC(t)	LabA(h1)	LabA(h2)	LabB(h)	LabC(h)
LabA(t1)	\	**0.16**	**0.13**	**0.22***	**0.19***	**0.19***	**0.14**	**0.25*$_{Max}$**
LabA(t2)	0.84	\	**0.09$_{Min}$**	**0.20***	**0.23***	**0.19***	**0.14**	**0.25*$_{Max}$**
LabB(t)	0.87	0.91$_{Max}$	\	**0.19***	**0.22***	**0.18***	**0.11**	**0.24***
LabC(t)	0.78*	0.80*	0.81*	\	**0.19**	**0.13**	**0.12***	**0.15**
LabA(h1)	0.81*	0.77*	0.78*	0.81	\	**0.20**	**0.19**	**0.22***
LabA(h2)	0.81*	0.81*	0.82*	0.87	0.80	\	**0.11***	**0.14**
LabB(h)	0.86	0.86	0.89	0.88*	0.81	0.89*	\	**0.19***
LabC(h)	0.75*$_{Min}$	0.75*$_{Min}$	0.76*	0.85	0.78*	0.86	0.81*	\

* Significant McNemar's test ($P < 0.05$): significant difference of proportion of positive results between the two runs; thus serious disagreement

$_{Min}$ Minimum

$_{Max}$ Maximum

negative for ISAV. Assuming that the PGS is correct, the targeted prevalence in the submitted samples (~ 50%) was reached which was fortuitous since the estimates within each salmon group did not agree with the ones reported in McClure et al. (2004). Using the PGS as a reference standard, the alignment of test results from the eight runs (3 laboratories, 2 sample types, and duplicates in lab A) highlighted the differences among test results (Fig. 2.2).

The computed unrooted tree represents the relative position among the eight test runs (Fig. 2.3A &B). Except for lab C, all non-homogenized samples were grouped together and formed a cluster supported by a low bootstrap value (61%). Within the cluster, lab A (duplicate 2) and lab B were the closest and associated with a high bootstrap value (88%) as previously shown in the agreement matrix (Table 2.3). Except for lab B, all homogenized samples were grouped together, including the lab C non-homogenized sample, and formed a cluster not supported by a high bootstrap value (50%). Within the homogenates cluster, both lab C samples and the lab A homogenate (duplicate 2) were grouped based on a low bootstrap value (60%). The lab C homogenate and lab A homogenate (duplicate 2) group was not supported by a high bootstrap value (43%) which does not separate them from the lab C non-homogenate. The results of the homogenized sample from lab B were consistently different, and therefore separated, from the two main clusters.

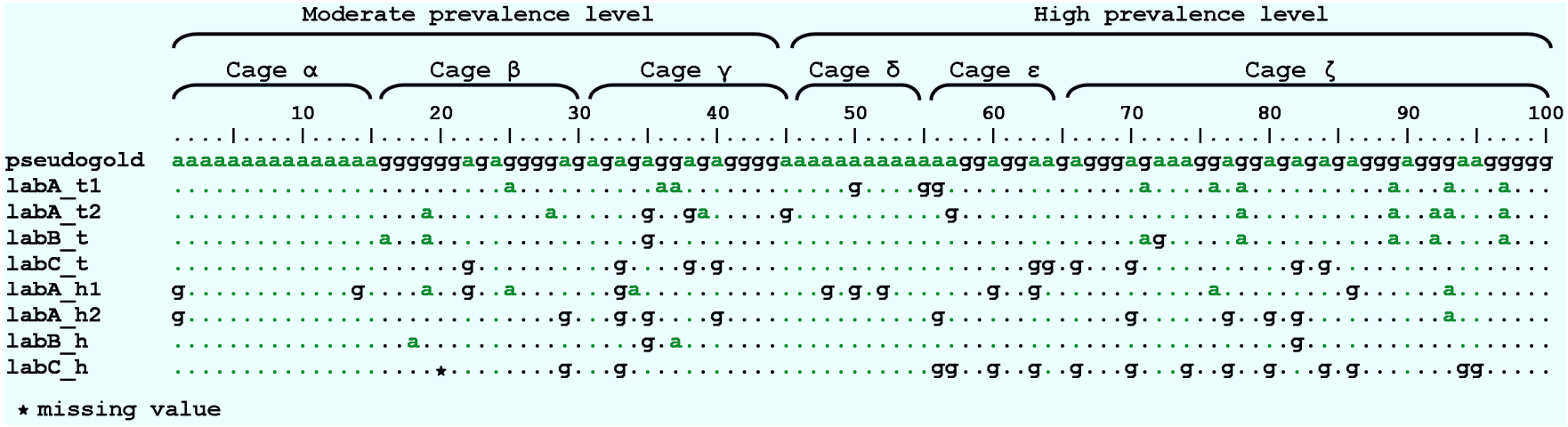


Fig. 2.2. Test result alignment: sampled salmon (in column) were clustered by cage origin and expected prevalence level population grouping (moderate level: apparently healthy fish from exposed cage; high level: mix of apparently healthy, mortality and moribund fish from infected cage). Negative was recoded as "adenine" (a); positive as "guanine" (g); and by column a dot (.) indicates same result as the first row. Greek letters: arbitrary cage number. (t1): non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample.

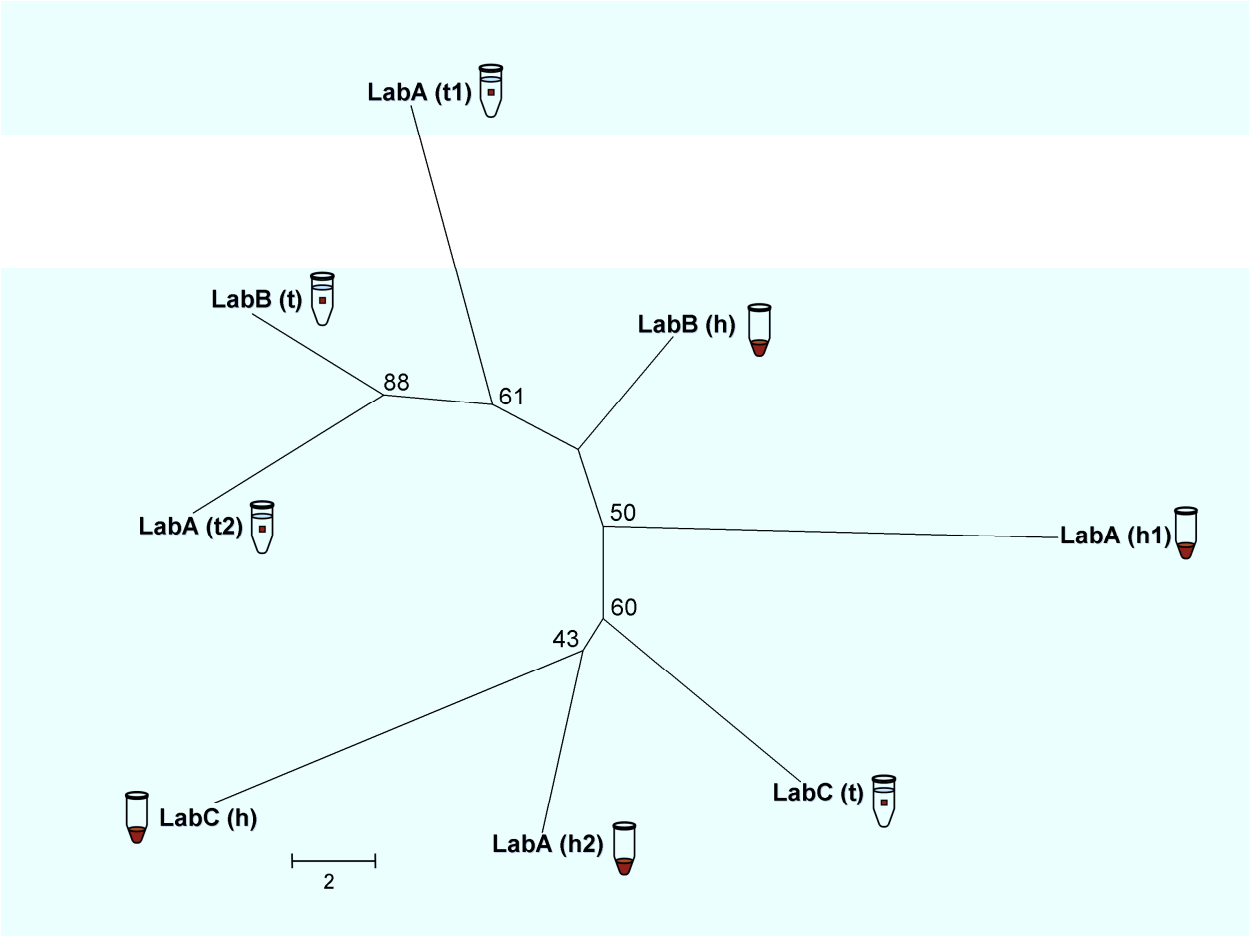


Fig. 2.3A . Star shaped unrooted phylogram representing agreement among test runs. The distance between two runs is visually assessed by the relative length of branches that connect them and are scaled based on the number of differing results out of the 100 samples tested. (t1): non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample.

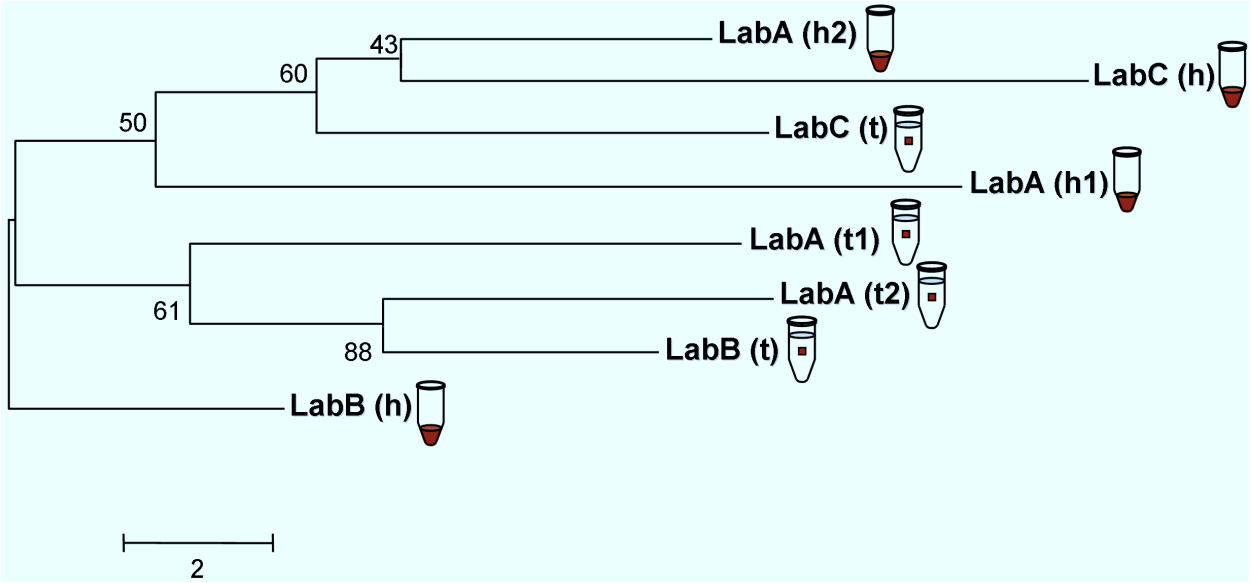


Fig. 2.3B . Tree shaped unrooted phylogram representing agreement among test runs. The distance between two runs is visually assessed by the relative length of branches that connect them and are scaled based on the number of differing results out of the 100 samples tested. (t1): non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample.

2.4. Discussion

2.4.1 Formal descriptive analysis of agreement

Although no minimum threshold has been set as suitable for validation of qualitative diagnostic tests, it is accepted that greater repeatability and reproducibility is preferred, provided that the McNemar's test is non-significant. Kappa values (κ) are usually used to compare test agreement beyond expected agreement (Dohoo et al., 2009). However, κ was estimated for all test runs at a same prevalence level of approximately 50%, and therefore, expected agreement due to chance would be consistent for all agreement estimates. Accordingly, we decided to base most of the discussion on Pa with little reference to κ values.

2.4.1.1 Repeatability

Based on the set of samples tested, the overall RT-PCR repeatability in the reference laboratory was approximately 81% (proportion of agreement). One could interpret this value as one in every set of five samples tested does not provide the same result when tested a second time (or 19% of the samples do not repeat the same result). When two results from the same individual disagree, based on a dichotomous outcome, one of the results has to be incorrect. Consequently, qualitative diagnostic tests that lack repeatability also lack accuracy. In this particular case, 9.5% (1/2 * 19%) of the combined results are either false positive or false negative.

Greater frequencies of false negative results imply decreased diagnostic sensitivity (DSe) that can be explained by several factors. The most likely reason for false negative results is a limited analytical sensitivity. Defined as the minimum threshold of detection, the analytical sensitivity of a RT-PCR depends on several method-specific factors including primer stringency (i.e. design of primer, nature and freshness of reagents), reaction preparation (i.e. ratio of target and primers), and thermocycling protocol (i.e. annealing temperature). It is possible that samples with a low concentration of target and/or a complex molecular matrix might not be detected if the primers do not bind to the target during the first cycles of the reaction. Thus, for infected samples with a concentration of RNA targeted close to the limit of detection, the target is sometimes detected or not, and the repeatability decreased. Among the 19% of non-repeatable results, some may be due to low concentration of target; hence the estimate of repeatability depends strongly on the nature of the sample tested. During routine surveillance, the pathogen load of screened apparently healthy individuals is likely to be low and the frequency of false negative results is expected to be high, whereas for confirmatory purpose, the pathogen load of clinically suspected individuals is likely to be high and the frequency of false negative results is expected to be low.

Greater frequencies of false positive results imply decreased diagnostic specificity (DSp) that can be explained by several factors. The most likely reason for false positive results with RT-PCR is cross-contamination (Wilson, 1997). Among the 19% of non-repeatable results, some may be due to contamination. Agreement among false positive results was not complete indicating that contamination was most likely random and not systematic. In theory, the probability of contamination should be associated with the

prevalence of infection in the sample pool tested. During routine surveillance, the prevalence of infected samples is usually low and the frequency of false positive results is expected to be low, whereas for confirmatory purpose, the prevalence of infected samples is likely high and the frequency of false positive results is expected to be more common.

As discussed previously (Begg, 1987; Greiner & Gardner, 2000), test operating characteristics depend strongly on the targeted population. Thus, the specific purpose and use of the diagnostic method must be clearly defined to reflect the assay's performance (OIE, 2008).

2.4.1.2 Reproducibility

The overall reproducibility of 82% proportion of agreement was estimated to be slightly higher than the overall repeatability of 81% (Table 2.1). In theory, it is expected to observe a larger variation in results between than within laboratories. Factors influencing the reproducibility include the ones influencing the repeatability plus factors differing among laboratory practices such as technician habit and training, equipment, facilities structure and organisation. For example, false positive or false negative results could arise from the subjective reading and interpretation of bands in electrophoresis gels. Laboratory technicians must choose the dichotomous result (i.e. positive or negative) according to the test protocol, their own training, and experience. Although mostly expected to influence the assay reproducibility, gel interpretation may also affect the repeatability due to human error and multiple laboratory staff.

A major source of decreased reproducibility is the discrepancy among laboratory operators in identifying a weak intensity band. Based on the subjective interpretation of

the operator, the sample will be retested or not. The subsequent interpretation in series of the repeated sample result (if the second test result is negative the sample is declared negative) may impact the agreement with laboratories that do not retest weak band samples. Retesting samples with weak bands aims to remove some false positive results and to increase DSp. This procedure may, however, also interpret as negative some truly low infected samples that are poorly repeatable. In this case, it would subsequently decrease DSe. If not standardized among laboratories, the retest of weak band samples and the interpretation of the second result would likely impact the reproducibility. Overall, it seems that these factors had little influence in this study since the overall agreements within and between laboratories were similar. However, these values reflect an average and do not reflect particular pairwise agreements.

Regardless of sample type, lab C had significantly higher proportions of positive results, suggesting serious disagreement with other laboratories (Table 2.1). This can be explained either by a higher DSe or by lower DSp in lab C. Assuming that the PGS is correct, lab C tended to have more false positive results (Fig. 2.2). Mostly with homogenates, lab B had overall a lower proportion of positive results. This can be explained by a lower DSe or a higher DSp in lab B. Assuming again that the PGS is correct, lab B seemed to have more true results than the two other laboratories (Fig. 2.2). There is evidence of variation in laboratory performance associated with low reproducibility.

In general, no international or standard guidelines are available to define acceptable levels of reproducibility. When two proportions of agreement are similar, the proportion of tests that agree on a positive result and the proportion of tests that agree on

a negative result can still differ. As an example, McNemar's test detected significant differences in proportions of positive results on homogenized tissue samples between lab B and C and not between lab A and B for identical Pa (Table 2.1). More sophisticated modelling, using multilevel logistic regression models, could alleviate the assumption of independent observations to simultaneously explore the effects of laboratories and homogenization on agreement levels (Chapter III).

2.4.2 Homogenization effect

The second objective of this study was to assess the effect of homogenization by comparing agreement between non-homogenized and homogenized samples. On average, homogenized samples had a higher proportion of positive results than non-homogenized samples (56.4% vs. 46.8%), which implies that homogenization impacted the test performances with either increased analytical sensitivity and DSe, decreased DSp, or both. Homogenized samples revealed slightly lower repeatability and reproducibility as non-homogenized samples (Table 2.1). We expected a strong improvement of agreement with homogenized sample as the supposedly heterogeneous distribution of ISAV particles in the salmon kidney was one suggested explanation for the low RT-PCR reproducibility in Nérette et al. (2005). Furthermore, significant symmetry and marginal homogeneity tests suggested a shift in testing pattern with homogenized samples. The marginal distribution of overall agreement revealed that the proportion of complete agreement (all test results agree for a given sample type) was higher with positive than with negative test results for homogenized samples while it was the opposite with non-homogenized

samples (Table 2.2). Although it was expected that higher complete agreement was observed for positive results with homogenized samples, a decrease of complete agreement for negative results was totally unexpected, in particular with the assumed dilution effect of homogenization (see below).

Reviewing the homogenization protocol and the fact that 12 fish with some positive results for homogenized samples were negatives for the four tests in non-homogenized samples (Table 2.2), it is feasible that cross-contamination occurred during homogenization. The use of pipette tips that lacked a filter might explain the potential false positives among the homogenates. Homogenization protocols, in particular of solid tissue, must be optimized and standardized in order to reach the maximal homogeneity in the sub aliquots. Even in a scenario of contamination, we would have expected all 4 homogenised aliquots to be contaminated. Random contamination with few viral RNA copies might thus explain a decreased repeatability or reproducibility with homogenized samples.

Repeatability and reproducibility estimates of tissue samples depend on the assumption that sub-samples from the same fish are identical and that the detection threshold of the assay is constant. Both can be either associated or independent. For example, in the initial phase of the infection, only clusters of low numbers of viral particles may be present in the salmon kidney to be tested. At this stage, homogenization would dilute already low levels of virus and produce more false negative results and lower agreement. Further, the progression of the infection would produce clusters of high numbers of viral particles as a result of viral replication. Homogenization would harmonize viral concentration among sub-samples at a detectable level despite dilution.

Finally, later stages of infection are expected to result in high numbers of viral particles throughout the organ. Homogenization would then provide little advantage since all tissue samples will contain high virus load. Although unrealistic according to the ISAV histopathology (Byrne et al., 1998), another scenario would be a spread of low numbers of viral particles throughout the organ. Homogenization would then provide limited benefit since each tissue sample would already have similar levels of particles. Agreement level would diminish mainly due to inconsistent detection of low virus load.

Overall, homogenization was of limited value in this precision evaluation; we suspect that occasional non-systematic error (cross-contamination) affected the samples' comparability and the agreement estimation. Tissue homogenization has diverse application (sample pooling, certified reference and control material, laboratory proficiency testing) and is greatly needed but an appropriate evaluation of its influence on test comparisons requires close monitoring and protocol optimization.

2.4.3 Novel descriptive analysis of agreement

2.4.3.1 Test result alignment

The approach offered in this study of using column (individual fish) and row (test run) to represent the test results similar to a genetic sequence alignment has not been previously published. This is a convenient and intuitive way for the reader and the investigator to screen and visually compare test results (Fig. 2.2), whereby each result is compared within a fish (column) to the first aligned test, in this case the PGS. No alignment algorithm is needed as each test result corresponds to a defined fish (or

column). From the alignment, it is possible to generate a matrix of pairwise comparisons among sequences, also called a distance or similarity matrix.

2.4.3.2 Test runs phylogram

The phylogram graphically represents the matrix of agreement and facilitates the visualisation of the relative position among test runs. Distance-based phylograms are generated from the matrix of pairwise genetic distances. A matrix of pairwise genetic distances is very comparable to a matrix of tests disagreement (1-Pa) (Table 2.2). Due to variable pressure of evolutionary changes, it is common in phylogeny to correct the estimates of genetic distance for multiple events per site (Van de Peer and Salemi, 2003). Since the probability that test results will be first positive then negative then positive again is extremely low, an evolutionary correction in the distance computation was judged not necessary. Future development of this approach may benefit from incorporating different weights for results changing from negative to positive and from positive to negative. Indeed, depending on the diagnostic test method being assessed, the probability of a false positive result (e.g. contamination) might be higher than the probability of false negative result (e.g. target decay during transport). More knowledge on the assay performances is, however, required to implement this refinement.

The distance matrix obtained from the alignment was identical to the initially computed disagreement matrix. Distance-matrix based tree reconstruction differentiates methods that are character-based and non character-based (Van de Peer and Salemi, 2003). The reconstruction generated by this study used only two arbitrary characters (adenine and guanine for negative and positive result, respectively) with equal weights of

substitution, giving no value to the character chosen. The phylogram construction was, however, rerun with the 12 possible combinations of two nucleotides from the four existing (adenine, guanine, cytosine, thymine) and confirmed the non-importance of character used as expected (data not shown).

Also referred to as pairwise distance methods, non character-based methods include cluster or minimum evolution analyses (Van de Peer and Salemi, 2003). The latter was preferred to the former because cluster analysis only assumes constant evolution (existence of a molecular clock) and would position all test runs in the tree equidistantly from the baseline or root. The commonly used method to estimate the minimal evolution tree is the Neighbor-Joining (NJ) method (Saitou & Nei, 1987). We selected only pairwise deletion in cases of missing data to avoid losing all the information from a fish when only one test result was missing. The obtained tree is a unrooted phylogram scaled for distances (set as the number of differing results) among test runs (Fig. 2.3A &B).

Bootstrap analysis is commonly used to evaluate the robustness of nodes that support tree branches. The magnitude of the bootstrap values is intimately correlated to the numbers of variable sites (or fish in this instance) that are informative in the alignment. A variable site is informative if there are at least two different characters that are represented at least twice at the given site. All bootstrap values, except for one node, were lower than 70% (Fig. 2.3A &B). The low resolution of the tree suggests some caution in its interpretation. With only 100 fish, the number of fish that discriminate the test runs might be limited and a higher number of salmon might provide a better tree

resolution. However, poor tree resolution will also be expected when test runs greatly agree (high consistency and precision).

The obtained phylogram illustrates the relative agreement among test runs (Fig. 2.3A &B). The distance between two runs is visually assessed by the relative length of branches that connect them. Non-homogenized samples were clearly clustered and showed some testing consistency; although lab C was separated, confirming poorer reproducibility. Within this cluster, non-homogenized samples of lab A (duplicate 2) and lab B were grouped separately from lab A (duplicate 1) which supported previous observations of similar repeatability and reproducibility on non-homogenized samples.

The cluster of homogenates, excluding lab B but including non-homogenized lab C, was poorly supported by bootstrapping (50%). This weak separation presumed a tendency of homogenized samples to test differently. The wide distribution of homogenates in the tree, however, supported a serious inconsistency in the testing pattern compared to non-homogenized samples. The homogenization protocol appeared to be inadequately refined or standardized to harmonize the testing pattern. Lab C revealed a distinct testing pattern with a reasonable repeatability regardless of the sample type and more closely resembling homogenized samples. However, lab C clearly decreased the overall assay reproducibility and must standardize its testing procedure to be comparable to the other laboratories.

The distance-matrix based tree reconstruction approach helps the investigator and the reader to visualize the relative proximity among test runs and to understand the distinctive testing patterns reflected by each of them.

2.5. Conclusion

Utilisation of basic phylogenetic reconstruction techniques provides a convenient and descriptive method to compare and assess agreement among test runs. The interpretation and validation of repeatability and reproducibility estimates, particularly using natural field samples, are complicated by the fact that no international standards and guidelines are established. Until guidelines are provided, we recommend considering as evidence of acceptable agreements results that show (i) fairly large κ estimates with (ii) a fairly narrow confidence interval obtained from (iii) a medium range prevalence, and (iv) conditional on a non-significant McNemar's test. Repeatability and reproducibility levels and the associated test accuracy appear to vary strongly with the intended use of assay. Appropriate assessment of consistency of test performance is critical to the interpretation of surveillance and control results and requires further development to model agreement across a range of population covariates (e.g. infection prevalences, infection stages) (Chapter III).

2.6. References

Agresti, A., 2002. Categorical data analysis. New York: Wiley.

Anonymous, 2000. ISA hits the Faroes. Fish Farming International. 27, 47.

Begg, C.B., 1987. Biases in the assessment of diagnostic tests. Stat. Med. 6, 411-423.

Bouchard, D.A., Brockway, K., Giray, C., Keleher, W., Merrill, P.L., 2001. First report of infectious salmon anaemia (ISA) in the United States. Bull. Eur. Assoc. Fish. Pathol. 21, 86-88.

Byrne, P.J., MacPhee, D.D., Ostland, V.E., Johnson, G., Ferguson, H.W., 1998. Haemorrhagic kidney syndrome of Atlantic salmon, *Salmo salar* L. J. Fish Dis. 21, 81-91.

Cleophas, T.J., Droogendijk, J., van Ouwerkerk, B.M., 2008. Validating diagnostic tests, correct and incorrect methods, new developments. Curr. Clin. Pharmacol. 3, 70-76.

Dohoo, I., Martin, W., Stryhn, H. (Eds.), 2009. Veterinary Epidemiologic Research. AVC Inc., Charlottetown, Canada.

Felsenstein, J., 1985. Confidence-limits on phylogenies - an approach using the Bootstrap. Evolution 39, 783-791.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical Methods for Rates and Proportions, 3nd Edition. Wiley, New York, USA.

Godoy, M.G., Aedo, A., Kibenge, M.J., Groman, D.B., Yason, C.V., Grothusen, H., Lisperguer, A., Calbucura, M., Avendaño, F., Imilán, M., Jarpa, M., Kibenge, F.S., 2008. First detection, isolation and molecular characterization of infectious salmon anaemia virus associated with clinical disease in farmed Atlantic salmon (*Salmo salar*) in Chile. BMC Vet. Res. 4, 28.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 42, 2-22.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids Symp. Ser. 41, 95-98.

ISO International Standard 5725-1, 1994. Accuracy (trueness and precision) of measurement methods and results. Part 1: General principles and definition. International Organisation for Standardisation (ISO), ISO Central Secretariat, 1 rue de Varembé, Case Postale 56, CH - 1211, Geneva 20, Switzerland.

Lee, J., Fung, K.P., 1993. Confidence interval of the kappa coefficient by bootstrap resampling. Psychiatry Res. 49, 97-98.

McClure, C.A., Hammell, K.L., Dohoo, I.R., Nerette, P., Hawkins, L.J., 2004. Assessment of infectious salmon anaemia virus prevalence for different groups of farmed Atlantic salmon, *Salmo salar* L., in New Brunswick. J. Fish. Dis. 27, 375-383.

Mullins, J.E., Groman, D., Wadowska, D., 1998. Infectious salmon anaemia in salt water Atlantic salmon (*Salmo salar* L) in New Brunswick, Canada. Bull. Eur. Assoc. Fish. Pathol. 18, 110-114.

Nérette, P., Dohoo, I., Hammell, L., Gagné, N., Barbash, P., MacLean, S., Yason, C., 2005. Estimation of the repeatability and reproducibility of three tests for infectious salmon anaemia virus. J. Fish. Dis. 28, 101-110.

Nérette, P., Stryhn, H., Dohoo, I., Hammell, L., 2008. Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic test. Prev. Vet. Med. 85, 207-225.

Office International des Epizooties, 2008. OIE Quality Standard and Guidelines for Veterinary Laboratories: Infectious Diseases. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 70pp.

Office International des Epizooties, 2009. OIE Aquatic Animal Health Code. 12th Edition. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 99-104.

Ransohoff, D.F., Feinstein, A.R., 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N. Engl. J. Med. 17, 926-930.

Rodger, H.D., Turnbull, T., Muir, F., Millar, S., Richards, R., 1998. Infectious salmon anaemia (ISA) in United Kingdom. Bull. Eur. Assoc. Fish. Pathol. 18, 115-116.

Saitou, N., Nei, M., 1987. The neighbor-joining method: anew method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406-425.

Stata Base Reference Manual, 2007. Volume 2, Q-Z, Release 10. A Stata Press Publication, Stata Corporation LP, College Station, Texas, USA, 536pp.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596-1599.

Thorud, K.E., Djupvik, H.O., 1988. Infectious salmon anaemia in Atlantic salmon (*Salmo salar* L). Bull. Eur. Assoc. Fish. Pathol. 8, 109-111.

Van de Peer, Y., Salemi, M., 2003. Phylogeny inference based on distance methods. In: Salemi, M., Vandamme, A.M. (Eds.), Phylogenetic Handbook: a practical approach to DNA and protein phylogeny. Cambridge University Press, Cambridge, pp. 101-136.

Vandamme, A.M., 2003. Basics concept of molecular evolution. In: Salemi, M., Vandamme, A.M. (Eds.), Phylogenetic Handbook: a practical approach to DNA and protein phylogeny. Cambridge University Press, Cambridge, pp. 1-23.

Wilson, I.G., 1997. Inhibition and facilitation of nucleic acid amplification. Appl. Environ. Microbiol. 63, 3741-3751.

Chapter III: A MODELLING APPROACH TO PREDICT THE VARIATION OF REPEATABILITY AND REPRODUCIBILITY OF AN INFECTIOUS SALMON ANAEMIA VIRUS RT-PCR ASSAY ACROSS INFECTION PREVALENCES AND INFECTION STAGES

Abstract

Traditional assessment of precision of binary outcome diagnostic tests focuses on descriptive estimates of agreement from a particular pool of studied samples. However, agreement for binary tests is intrinsically associated with precision (random error) and accuracy of the assay (systematic error). When two test results disagree, one is correct and the other incorrect. Assay operating characteristics are potentially strongly influenced by the test population and laboratory covariate factors that may result in agreement variations. Using test result information from a previous descriptive study on agreement within and between laboratories (repeatability and reproducibility, respectively) of an RT-PCR assay for infectious salmon anaemia virus (ISAV), the influence of submission factors (tissue homogenization, infection prevalence and pathogen level) on agreement was further investigated. Multilevel logistic regression models were constructed separately for non-, low- or high-infected salmon (classified using a study pseudogold standard) to predict probabilities of testing positive under different testing conditions. For given prevalences and infection stages of infected fish, agreement and kappa values were computed from the predicted values using probability formulae and category weighting. Repeatability and reproducibility varied greatly with prevalence, and the influence of infection stages on agreement was lowered by homogenization which supported a heterogeneous distribution of ISAV in early infected salmon kidney. This predictive approach provided a better expectation of assay agreement and increased the capacity to apply and extrapolate estimates of repeatability and reproducibility to other circumstances of use.

3.1 Introduction

3.1.1 Diagnostic precision

Evaluation of diagnostic assay precision and accuracy are critical components of the multi-stage validation procedure recommended by the World Organisation for Animal Health (OIE) to certify methods that either diagnose disease or detect the presence of an associated pathogen (i.e. infection) (OIE, 2008). Assay precision is conventionally assessed by estimating repeatability and reproducibility. According to International Standards definitions (IS0 5725-1, 1994), the repeatability is defined as the variation in test results obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time. Reproducibility is defined as the variation in test results obtained with the same method on identical test items in different laboratories with different operators using different equipment. For a dichotomous assay (i.e. binary outcome), variation in test results is estimated by the agreement between two test runs. We define here "test run" or "run" as a set of results obtained using the same method under defined testing conditions that may differ within a laboratory for repeatability, and between laboratories for reproducibility. Traditionally, estimates of within- and between-laboratory agreement are expressed as proportion of agreement (proportion of tests results that agree) (Pa) or as Cohen's Kappa values (κ) (Dohoo et al., 2009).

3.1.2 Diagnostic agreement relativity

Albeit agreement is commonly assumed to be a measure of test precision, the lack of result consistency is in fact a combination of imprecision (random error) and inaccuracy (systematic error) (Van der Bruel et al., 2007). When test results from two identical samples do not agree, one has to be correct and the other not. Alternatively, when two test results agree on a positive, they are either both right (substantial diagnostic sensitivity, DSe) or both wrong (poor diagnostic specificity, DSp) depending on the true status of the sample. Equally, when two test results agree on a negative, they are both correct (substantial DSp) or both incorrect (poor DSe). Agreement therefore depends directly on DSe and DSp. Interestingly, strong agreement in infected/diseased samples results from either both very high DSe (both results positive) or very low DSe (both results negative). Reciprocally, strong agreement in non-infected/non-diseased samples results from either both very high DSp (both results negative) or very low DSp (both results positive). Therefore, at the animal level, agreement depends primarily on the magnitude of DSe and DSp, and also on their stability. Agreement relies also on the degree of dependence between sub-samples collected from a same individual. Cleophas et al. (2008) prefer to report assay precision using predictive intervals of DSe, DSp or overall accuracy. To match international requirements for test validation (OIE, 2008), we restricted the evaluation of test precision to the estimation of repeatability and reproducibility.

Even if DSe and DSp were appreciable and stable, estimates of agreement may also vary greatly with prevalence of infected/diseased animals in the tested population

(population level). If agreements within infected/diseased and non-infected/non-diseased individuals differ, overall agreement reflects the average of both status-specific agreements weighted on the proportion of each status in the tested sample pool (i.e. prevalence). Consequently, overall agreement estimates may vary substantially across prevalences of infected/diseased individuals. To limit extrapolation error from specific population estimates, the Standards for Reporting of Diagnostic Accuracy (STARD) recommended reporting variation of diagnostic performances across populations (Bossuyt et al., 2003). Therefore, evaluation of test precision should include infection status-specific estimates and a description of agreement variation across a range of infection/disease prevalences. Assuming that agreement for each infection/disease category are known and constant, it is possible to compute (and predict) agreement across a range of prevalences using simple category weighting, as suggested by Björk et al. (2009).

The operating characteristics of an assay may vary greatly with testing conditions, such as sample preparation or processing laboratory. Collection technique, storage, buffer solution, and homogenization are sampling factors that are associated with test accuracy and might change the agreement within an infection/disease category. Intrinsic laboratory factors such as instrument, operator, and facility design are also associated with test accuracy and might also influence agreement within an infection/disease category. It is therefore essential to refine the prediction of agreement by investigating variations of agreement associated with submission covariate factors.

3.1.3 ISAV RT-PCR repeatability and reproducibility

Infectious Salmon Anaemia virus (ISAV) (family Orthomyxovirus, genus *Isavirus)* is the causative agent of an important infectious disease, particularly of Atlantic salmon (*Salmo salar* L.). Targeting endothelial cells of salmonids, the viral pathogen is found in Northern Europe (Thorud and Djupvik, 1988; Rodger et al., 1998; Anonymous, 2000), and in North and South America (Mullins et al., 1998, Bouchard et al., 2001, Godoy et al., 2008). Responsible for high cumulative mortality when present in salmon farms, ISAV is listed as a reportable aquatic disease by the OIE (OIE, 2009a), requiring intensive surveillance and control programs.

ISAV surveillance was identified as part of the Canadian National Aquatic Animal Health Program (NAAHP), and required that a recently developed Reverse-Transcriptase Polymerase Chain Reaction (RT-PCR) assay (Gagné et al., data unpublished) be evaluated, validated and certified. The targeted use of the assay in Atlantic salmon aquaculture and wild fish samples is to demonstrate freedom of infection (not disease) in a sea-cage, farm site, bay, or region in Canada. The assay is able to detect the presence of a portion of RNA from the 8th segment of the viral genome in a kidney tissue sample which is assumed to be correlated with the presence (or absence) of "active" viral particles in the fish. Following the bench validation phase, the first stage of this field evaluation focuses on the estimation of the assay's repeatability and reproducibility.

Two previous studies reported estimates for ISAV RT-PCR repeatability and reproducibility using different protocols and evaluation designs. The first study,

published by Nérette et al. (2005), examined ISAV RT-PCR agreement within and among three different laboratories and revealed poor reproducibility and substantial differences in repeatability. The authors, however, described factors associated with the study design that may have biased the comparability of test results. First, different kidney tissue samples from the same fish were submitted for detection in the different laboratories. These samples were though to be hypothetically heterogeneous for the distribution of viral particles in the organ. If the sub-samples differed, then they may have compromised the requirement of "identical test items" for agreement evaluation. Second, different amplification protocols (i.e. different set of primers, reagents and methodology) were used in different laboratories and may have compromised the requirement of using the "same test". Third, different result interpretations and confirmation protocols (i.e. sample with a weak gel band retested) may also have compromised the comparability of laboratories. In a subsequent study, presented in Chapter II, described the repeatability and reproducibility of a different ISAV RT-PCR assay in three different laboratories while specifically addressing the design limitations described in Nérette et al. (2005). Homogenized kidney sub-samples were submitted for testing in parallel to reduce the impact of within-tissue variability. In addition, all participating laboratories agreed to use the same detailed test protocol to detect ISAV. Lastly, gel band interpretation and retesting strategies were harmonized among the three laboratories. Regardless of the sample type (homogenized or non-homogenized), the descriptive analysis of the test results revealed only a moderate repeatability in the reference laboratory and a slightly lower overall laboratory reproducibility. It is believed that this was weakened by the strong disagreement of one participating laboratory compared to the other two

laboratories. Due to substantial study design and testing protocol discrepancies, the two study estimates can not and have not been directly compared.

In addition, both Nérette et al. (2005) and the Chapter II only reported overall agreement estimates specific to respective study sample pools. The Chapter II discussed the difficulty in generalizing their estimates due to the close relationship between agreement and accuracy and the need to explore agreement variation across population covariate factors such as prevalence. Chapter II further suggested that DSe, and associated agreement in infected fish, would be strongly dependent on pathogen load in the tested samples. Referred to as "spectrum" (Ransohoff and Feinstein, 1978) or "case-mix" (Begg, 1987), the varying proportions of infection stages associated with viral load would potentially influence agreement within infected individuals, regardless of the prevalence. Variation of agreement should also be investigated across a range of infection stages in infected fish.

3.1.4 Study objectives

Recognizing that repeatability and reproducibility are proportion-weighted averages of agreements specific to infection stage, the objective of this study was to predict test agreement across a range of infection stages and prevalences for ISAV in salmon. The modelling approach used also allowed us to assess the effect on agreement of submission factors, including sample preparation (i.e. homogenization) and testing laboratory.

3.2 Material and methods

3.2.1 Data, complementary testing, and a pseudogold standard

3.2.1.1 Data

The data used in this study were derived from a previous study on the descriptive analysis of repeatability and reproducibility of a recently designed RT-PCR for ISAV (Chapter II). Kidney samples from 100 Atlantic salmon, *Salmo salar*, were collected in six replicates from each fish. Duplicate coded tissue samples were tested by the reference laboratory (lab A) and single samples were tested by two other laboratories (lab B & C). The remaining two tissue samples were combined, homogenized and aliquoted into four coded microtubes. Two of these homogenates were tested in lab A and single homogenates were tested by the other two laboratories. Full descriptions of sample allocation and testing protocol are reported in Chapter II.

3.2.1.2 Complementary testing (Real-time Reverse-Transcriptase Polymerase Chain Reaction)

Two hundred RNA extracts, obtained from the duplicate tissue samples originally analyzed by conventional RT-PCR in the reference laboratory (lab A), were recoded, randomly ordered and tested again using the real-time version of the previous RT-PCR protocol. Reverse transcription was done with the High Capacity Reverse transcription kit

(Applied Biosystems). RNA (≤ 1000 ng/µl) and water were mixed, denatured at 95°C for 5 min and transferred to ice. The enzyme mix containing 2 µl of random primers was added to the tubes, for a total of 20 µl per tube, and incubated according to the manufacturer's instructions; 20 µl of H_2O was added after the reaction. Real-time PCR was performed with the Taqman Universal PCR Master Mix (Applied Biosystems), on a Mx3000P thermocycler (Stratagene). PCR was done in 25 µl volumes comprising 2 µl of cDNA, 0.48 µM of each primers, and 0.2 µM of probe. Primers were 404F_ISA8: 5' tgg gca atg gtg tat ggt atg a-3' and RA3 (583R)-ISA8: 5' gaa gtc gat gaa ctg cag cga-3'; the FAM labelled probe (491_ ISA8) sequence was 6-FAM cag gat gca gat gta tgc-MGB (quencher). Cycling conditions consisted of an initial hold at 50 °C for 2 min, then 95 °C for 10 min, and 45 cycles at 95 °C – 30 sec, 60 °C – 30 sec, 72 °C – 30 sec, with fluorescence reading at the end of each cycle. For quality control, extraction blanks (no samples) were included every 15th tube during extractions, and blanks (water) were added for the reverse transcription and the PCR steps; positive controls were included during the PCR step. Any obtained cycle threshold (Ct) values were reported as positive results; all blanks and negative controls yielded no Ct values.

3.2.1.3 Pseudogold standard

ISAV replication was assumed to be sigmoid across time with three successive infection stages characterized by increasing viral load: low, intermediate and high infection. The duration of the intermediate stage is limited and therefore under-represented in an infected population. Consequently, we only considered low- or high-

infection stages in this study. Each of the 100 salmon was thereafter classified as non-, low- or high-infected according to an established pseudogold standard (PGS). Salmon were first classified as infected and non-infected using the criteria of Chapter II based on the combination of the six previous results from the conventional RT-PCR (excluding duplicate results in lab A). A fish was classified as "non-infected" (NI) if it yielded three or less positive tests out of the six (≤ 3/6). Alternatively, a fish was classified as "infected" if it yielded more than three positive tests out of the six (> 3/6). "Infected" fish were further dichotomized into low- or high-infected using additional real-time RT-PCR duplicated results. Obtained Ct values from the tested samples ranged from 17.41 to 42.98, and by extrapolation, we considered that Ct values from any ISAV infected salmon would universally range from 15 to 45 cycles. Thereafter, we arbitrarily selected the range mid-point (30 cycles) as the cutoff to sub-classify infected fish as low- or high-infected. "Low-infected" (LI) fish were infected salmon with the lowest obtained Ct value equal to or higher than 30 cycles. "High-infected" (HI) fish were infected salmon with the lowest obtained Ct value lower than 30 cycles. Six salmon classified as NI had one duplicate sample test positive to real-time RT-PCR; and 2 salmon classified as infected had both duplicate samples test negative to real-time RT-PCR (no Ct) and were thus classified as LI. Test results from conventional and real-time RT-PCR were aligned for visual comparison using PGS as the reference using the approach described in Chapter II (Fig. 3.1).

Two alternative PGS definitions were further developed to investigate the influence of the classification criteria on the analysis outcome. The first alternative PGS (Strict-PGS) used more restrictive criteria to classify fish as infected such that a salmon

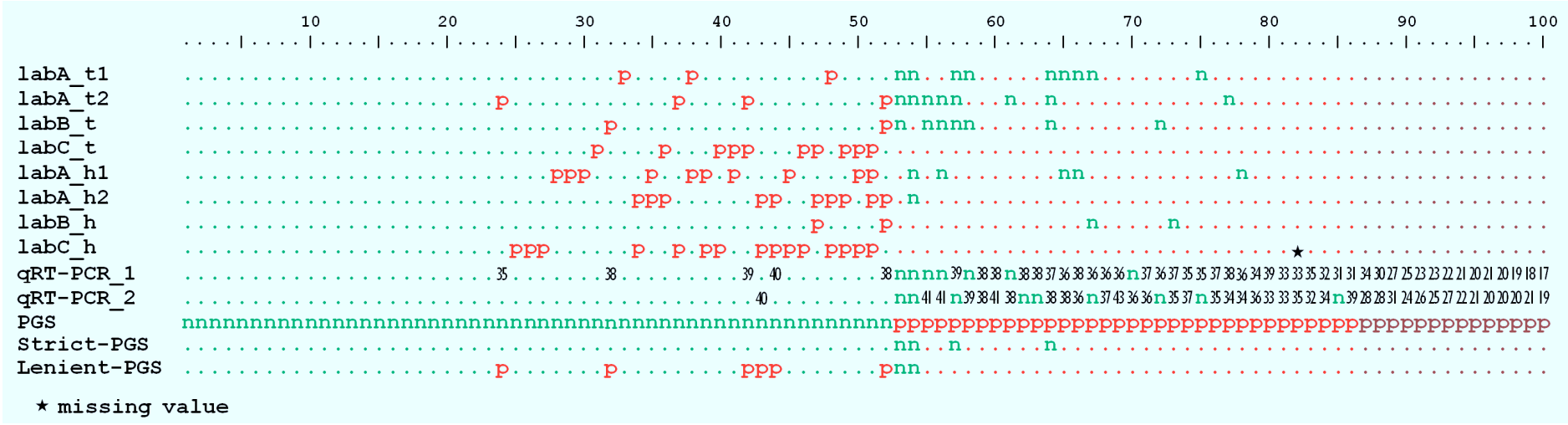


Fig. 3.1. Test result alignment: sampled salmon (in columns) **were clustered by infection stage according to the pseudogold standard** (PGS). Results are coded as "n" for negative or "p" for positive; and a dot (.) indicates same result as the PGS by column. Non-infected salmon were defined as fish with 3 or less positive test results out of the 6 conventional RT-PCR assays run from the two sample types (i.e. (t): non-homogenized sample, (h): homogenized sample) in the 3 participating laboratories (lab A, B & C) without including duplicates in lab A (i.e. (t2) & (h2)). Within non-infected salmon, fish were ordered by increasing number of positive test results. Within infected salmon (fish with more than 3 positive test results), fish were classified and ordered as low-infected (light "p") or high-infected (dark "p") according to the duplicated real-time RT-PCR (qRT-PCR) results reported as cycle threshold (Ct) values. Low-infected were defined as fish with the minimum yield Ctvalue equal or above 30; while high-infected were fish with the minimum Ct value below 30. Strict and Lenient-PGS are alternative standards used for comparison, whereby Strict-PGS classified non-infected fish as 4 or less positive conventional RT-PCR (out of 6), and Lenient-PGS classified non-infected fish as negative on both qRT-PCR duplicates. The cutoff for low- and high-infected fish remains constant across the 3 pseudogold standards.

had to be positive on the two types of samples in at least two laboratories (i.e. at least five positive results out of the six conventional RT-PCR runs) (Fig. 3.1). The empiric cut-point to sub-classify infected fish was conserved for both additional PGS definitions. The second alternative PGS (Lenient-PGS) used less restrictive criteria to classify fish as infected and only used duplicated results from real-time RT-PCR. To be classified as infected, a salmon had to be positive on at least one real-time RT-PCR run (Fig. 3.1).

3.2.2 Multilevel logistic models

3.2.2.1 Data hierarchy and model construction

For analysis, the dataset was split into three subsets according to the PGS status and separate multilevel logistic models were conducted for each infection category (NI, LI, HI). MLwiN software v.2.11 (Rasbash et al., 2009) was used for model building and as a transfer platform. Across all laboratories and sample types, all nominated HI individuals tested positive (Fig. 3.1). Therefore, no model was developed for HI fish since their probability to test positive was 1 regardless of the assay conditions. The original hierarchical structure included 3 levels: "fish" (100); "samples within fish" (five in total: four separate tissues and one homogenate); and "tests within sample" (four for homogenized sample and one for non-homogenized sample) (Fig. 3.2A). Preliminary model construction did not reveal any variation at the "sample" level for both NI and LI fish after accounting for homogenization. Therefore, finalized models only included two levels: "fish" and "test results within fish" (eight tests for each fish) (Fig. 3.2B).

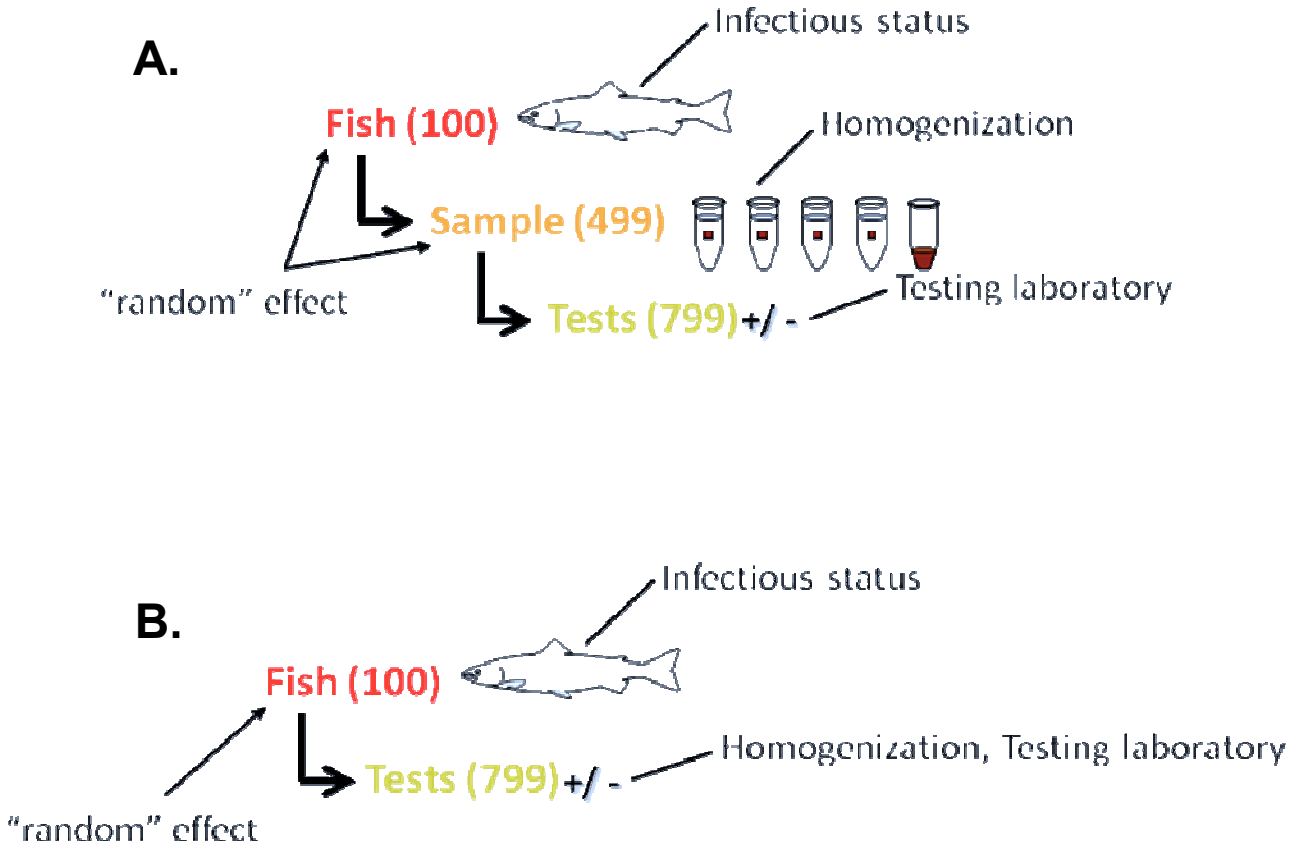


Fig. 3.2. Hierarchical structure of the dataset: 3-level structure (A); 2-level structure (B).

Multilevel logistic regression models included "fish" as a random effect, and homogenization (Hom), laboratories (Lab) and their interaction as fixed effects. Identical models were initially built for NI and LI fish:

$$logit\ (Pr(test{:}pos)) = \beta_0 + \beta_1\ Hom + \beta_2\ Lab\ B + \beta_3\ Lab\ C + \beta_4\ Hom{*}Lab\ B + \beta_5\ Hom{*}Lab\ C + u_{fish} \quad (1)$$

where u_{fish} is the random effect from N(0, σ^2_{fish}), a normal distribution with mean 0 and variance σ^2_{fish}. Regardless of the sample type, all nominated LI individuals tested positive in lab C (Fig. 3.1). Since the probability to test positive in lab C was 1, lab C results were removed from the LI dataset and the model was developed without the lab C fixed effect. In the model from Eq. (1), test results in different laboratories were conditionally independent given the random effect (and true sample status); this is a standard way of incorporating (conditional) test independence (e.g. Yang and Becker, 1997)

3.2.2.2 Bayesian analysis

For reasons detailed below (see section 3.2.2.3), predictions from model (1) were based on the posterior distributions obtained in a Bayesian estimation. From MLwiN, the models were transferred to WinBUGS v.1.4.3 software (Spiegelhalter et al., 2003) for analysis. Markov-chain Monte Carlo models were run using Gaussian priors for fixed effects regression coefficients (mean =0, variance = 1.E-6) and a uniform prior (0,100) for σ^2_{fish}, a burn-in period of 10,000 iterations, and a run of 1,000,000 iterations with

thinning sampling every 100 iterations for a total of 10,000 posterior samples. Regardless of their significance, all fixed factors were kept in the models as the main modelling objective was prediction. Three Markov chains were run in parallel with different sets of initial values for model parameters to conduct the Gelman-Rubin convergence diagnostic (Brooks and Gelman, 1998) but no influence on posterior estimates was detected. Model convergence was assessed visually with informal examinations (Gelman, 1996) involving the convergence diagnostics based on quantiles (Raftery-Lewis) and means (Brook-Draper) provided by MLwiN. Poor autocorrelation along the Gibbs sampler chains was confirmed visually.

3.2.3 Agreement computation

A multilevel logistic regression model, Eq. (1), yields parameters with subject-specific (SS) interpretation (Zeger et al., 1988). As test and agreement characteristics should be interpreted across the population of samples, a conversion to population-averaged (PA) estimates is required (see McClure et al. (2005) for a discussion of the context related to test characteristics). Test characteristics such as DSe and DSp may be calculated directly based on predicted probabilities from Eq. (1), after an approximate conversion of parameters to a PA interpretation (Dohoo et al., 2009). Agreement and kappa statistics do not permit a closed formula calculation but involve integrals over the random effects distribution. Therefore, we used an integral approximation by averaging over randomly sampled random-effects terms (i.e. a Monte Carlo approximation to the integral) (Smyth, 2005). All test and agreement characteristics were calculated by

averaging over random effects sampled for each of the 10,000 posterior samples, using the detailed formulae outlined below. The advantage of this approach is that it incorporates uncertainty about the parameter estimates by using values across the posterior distribution. Subsequent calculations were carried out in standard spreadsheet and statistical software.

3.2.3.1 Estimated agreement

We termed repeatability (*r*) the estimated agreement between duplicate samples in the reference laboratory (lab A). We termed reproducibility (*R*) the computed average of estimated agreements of duplicate samples between labs A and B, A and C, and B and C. Generally, estimated agreement (*Est. Agr.*) between samples analyzed at two defined testing conditions c1 and c2 (which could involve the same or different labs) was computed by the formula:

$$Est.\ Agr.(c1,c2|I,u) = Pr(pos,c1|I,u) * Pr(pos,c2|I,u) + Pr(neg,c1|I,u) * Pr(neg,c2|I,u) = Pr(pos,c1|I,u) * Pr(pos,c2|I,u) + (1 - Pr(pos,c1|I,u)) * (1 - Pr(pos,c2|I,u)) \quad (2)$$

where *I* refers to the infection stage (NI, LI, HI) and *u* to a particular posterior random sample from respective models. Conditional probabilities are computed from the linear predictors (β) on the logit scale and the inverse logit transformation ($\text{logit}^{-1}(\beta) = 1 / (1 + e^{-\beta})$) (Dohoo et al., 2009). The dependence on the random effect is addressed by averaging the estimates over the 10,000 posterior samples. For instance, the PA estimate

of agreement between condition c1 and c2 in the infection category *I* is obtained as follows:

$$Est.\ Agr.(c1,c2|I) = (^{1}/_{10,000})\ \Sigma_i\ [\ Pr_i(pos,c1|I,u_i) * Pr_i(pos,c2|I,u_i) + Pr_i(neg,c1|I,u_i) * Pr_i(neg,c2|I,u_i)] \quad (3)$$

where u_i is the random value sampled from $N(0, \sigma^2_i)$; and σ^2_i and β_i inherent in Pr_i are the parameters of the i^{th} sample of the posterior . For a mixture population with a prevalence (θ) and a proportion (π) of LI among all infected fish, the conditional estimated agreement was computed as:

$$Est.\ Agr.(\ c1,c2|\theta,\ \pi,\ u) = \theta * (\pi * Est.\ Agr.(LI,u) + (1-\pi) * Est.\ Agr.(HI)) + (1-\theta) * Est.\ Agr.(NI,u) \quad (4)$$

where *Est. Agr.(LI,u)*and *Est. Agr.(NI,u)* are estimated agreements in LI and NI, respectively. Note that for simplicity of notation, we used the same symbol *u* to denote both random effects, although the values for LI and HI models were sampled from different distributions (hence different iteration) and randomly combined. For HI fish, *Est. Agr.(HI)* = 1, as described above. Similar to Eq. (3), the PA estimates were computed by averaging over the 10,000 conditional estimates.

3.2.3.2 Chance agreement and Cohen's kappa

Cohen's kappa (κ) value involves the agreement beyond chance, i.e. the expected agreement computed if no correlation between test results is assumed (Dohoo et al., 2009). We first describe how the model (1) was used to compute the chance agreement. In the absence of information about other test results, the probability to test positive (or negative) under a given testing condition depends on the corresponding test characteristics (DSe and DSp) as well as on the prevalence (θ) and the proportion (π) of LI fish:

$$Pr(pos|\theta, \pi, u) = \theta * (\pi * DSe(LI,u) + (1- \pi) * DSe(HI)) + (1 - \theta) * (1 - DSp(u)) \quad (5)$$

$$Pr(neg|\theta, \pi, u) = \theta * (\pi * (1 - DSe(LI,u)) + (1- \pi) * (1 - DSe(HI))) + (1 - \theta) * DSp(u) \quad (6)$$

where u refers to a particular posterior random sample from respective models. For HI fish, DSe(HI) = 1, as described above. Similar to Eq. (3), the PA estimates of DSe(LI) and DSp for a specific condition (c) were computed by averaging the predicted probabilities of testing positive in LI fish (DSe(LI)) or negative in NI fish (DSp) over the 10,000 posterior samples.

Thereafter, the chance agreement between two testing conditions (c1 and c2) was computed as:

$$Ch.\ Agr.(c1,c2|\theta, \pi, u) = Pr(pos,c1|\theta, \pi, u) * Pr(pos,c2|\theta, \pi, u) + Pr(neg,c1|\theta, \pi, u) * Pr(neg,c2|\theta, \pi, u) \quad (7)$$

where *Pr (pos,c1|θ, π)* refers to the probability of testing positive under testing condition c1. Finally, Cohen's kappa (κ) between testing conditions (c1 and c2) was computed by averaging the Kappa calculation formulae as follows:

$$\kappa (c1,c2|\theta, \pi, u) = (\text{Est. Agr. } (c1,c2|\theta, \pi, u) - \text{Ch. Agr. } (c1,c2|\theta, \pi, u)) / (1 - \text{Ch. Agr. } (c1,c2|\theta, \pi, u)) \quad (8)$$

Similar to Eq. (3), the PA estimates was computed by averaging over the 10,000 conditional estimates of κ.

3.2.4 Alternative approach to modelling

A convenient alternative approach was used to directly estimate probabilities to test positive (or negative) in each infection category (NI, LI, HI) as defined by the PGS. For instance, from Fig. 3.1, positive test results for NI fish in lab A relative to replicated non-homogenized samples were counted (7 positives out of 104 results) and the corresponding probability estimated (6.7%). Thereafter agreements were calculated by inserting the respective probabilities in Eqs. (2), (5) and (6). Using simple descriptive statistics, these agreement estimates were referred as *descriptive* estimates. Descriptive estimations were used when assessing the impact of the three PGS definitions.

3.2.5 Agreement graphical representation

Three sets of graphics were generated for each agreement type (i.e. estimated, chance, and κ) including within- and between-laboratory comparisons for homogenized and non-homogenized samples. Each graph represents agreement across a range of infection prevalences (from 0 to 100%) and for three proportions of LI among all infected fish (0, 50 and 100%). In addition, the agreement profile corresponding to the supposed proportions of LI and HI fish among all infected fish (71% LI, 29% HI) was added to be compared to the actual observed value of agreement. Minimum and maximum limits of the predictive interval (95%) of estimated, chance and κ agreements were calculated using the 2.5% percentile and 97.5% percentile (not the average) of Eqs. (2), (7) and (8) for each prevalence and submission condition, regardless of the proportion of LI.

3.3 Results

3.3.1 Pseudogold standard and observed values

Among the 100 tested salmon, the PGS identified 52 non-, 34 low- and 14 high-infected (NI, LI and HI, respectively) fish. Assuming the PGS is correct, the prevalence of infected fish was 48% and the proportion of LI fish among all infected was 71% (and 29% HI). Alternatively, the Strict-PGS identified 44% infected salmon among which 68% were LI and 32% were HI, while the Lenient-PGS identified 52% infected salmon

among which 73% were LI and 27% were HI. It was noteworthy that the proportion of LI among the infected fish did not change substantially with the three PGS versions (71%, 68% and 73%, respectively). Descriptive statistics of this dataset including *observed* agreement proportions (Pa) and kappa values (κ) were reported elsewhere (Chapter II). Observed *r* within the reference laboratory for non-homogenized and homogenized samples were 0.84 and 0.80, respectively, while the corresponding values of averaged *R* were 0.82 and 0.80. Assuming test independence, observed chance *r* within the reference laboratory for non-homogenized and homogenized samples were 0.51 and 0.50, respectively, while the corresponding values of averaged chance *R* were 0.50 and 0.50. The observed κ within the reference laboratory for non-homogenized and homogenized samples were 0.67 and 0.69, respectively; while the corresponding values of κ among the 3 participating laboratories were 0.64 and 0.60.

3.3.2 Multilevel logistic models

For each model, posterior distributions of regression coefficients (homogenization, testing laboratory and first-order interaction) and variance estimates at the fish level are numerically summarized in Table 3.1. A direct interpretation of back transformed model coefficients would be subject-specific (SS) (Dohoo et al., 2009), and therefore, for proper population-averaged (PA) interpretation, one can adjust regression coefficients using a conversion formula: $\beta_{PA} \approx \beta_{SS} / \sqrt{(1 + .346\ \sigma^2)}$ (McClure et al., 2005). PA adjusted estimates by conversion were reported in the sections below to facilitate the interpretation of model coefficients. For comparison, PA estimates were also

Table 3.1
Summary of posterior distributions of parameters estimated in subdatasets defined by a pseudogold standard (PGS).

Parameters	Non-Infected Fish Model		Low-Infected Fish Model		High-Infected Fish
Number of records	*416 test results used from 52 fish*		*204 (271*) test results from 34 fish*		*112 test results from 14 fish*
Random Effects	*Variances (σ^2)*	*C.I.(2.5%,97.5%)*	*Variances (σ^2)*	*C.I.(2.5%,97.5%)*	
Fish	**1.108**	**0.45, 1.85**	**1.906**	**0.90, 3.34**	na
Fixed Effects	*Coefficient*	*C.I.(2.5%,97.5%)*	*Coefficient*	*C.I.(2.5%,97.5%)*	
Constant (Lab A, tissue)	**-3.16**	**-4.22,-2.25**	**1.77**	**0.77, 3.08**	na
Homogenization	**1.38**	**0.42, 2.44**	**1.70**	**0.55, 2.95**	na
Lab B	-0.82	-2.90, 0.80	0.38	-0.81, 1.62	na
Lab C	**1.36**	**0.27, 2.54**	na	na	na
Homogenization x Lab B	-1.36	-3.95, 1.12	0.32	-1.81, 2.77	na
Homogenization x Lab C	-0.73	-2.16, 0.65	na	na	na

Bold: parameter significantly different from 0
*record including data from lab C
C.I.: credibility interval
na: not applicable since the probability to test positive is perfect (100%)

Table 3.2

Probabilities for infectious salmon anaemia virus RT-PCR to test negative or positive according to infection status, the sample type (homogenized or non-homogenized)**, the processing laboratory** (lab A, B or C) **and the infection stage** (non-, low-, high-infected)**.** Classification of fish in infection stages originally followed a preset pseudogold standard (PGS). For PGS, the probabilities were first computed as population average estimates from averaged model predicted values (Modelling) or by direct description of probabilities in each infection category (Descriptive). Descriptive probabilities were added using two alternative pseudogold classifications (i.e. Strict- & Lenient-PGS) to assess the impact of the pseudogold definition on prediction.

	Probability to test negative in Non-infected salmon						Probability to test positive in Low-infected salmon						Probability to test positive in High-infected salmon					
	Non-homogenized			Homogenized			Non-homogenized			Homogenized			Non-homogenized			Homogenized		
	lab A	lab B	lab C	lab A	lab B	lab C	lab A	lab B	lab C	lab A	lab B	lab C	lab A	lab B	lab C	lab A	lab B	lab C
Modelling PGS	0.930	0.960	0.806	0.806	0.958	0.712	0.747	0.788	1.00*	0.902	0.933	1.00*	1.00*	1.00*	1.00*	1.00*	1.00*	1.00*
Descriptive PGS	0.933	0.962	0.808	0.808	0.962	0.712	0.750	0.794	1.000	0.912	0.941	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Descriptive Strict-PGS	0.937	0.946	0.750	0.768	0.893	0.661	0.850	0.934	1.000	0.934	0.934	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Descriptive Lenient-PGS	0.958	0.98	0.771	0.702	0.917	0.688	0.711	0.737	0.868	0.816	0.816	0.895	1.000	1.000	1.000	1.000	1.000	1.000

* not modelled since all samples tested positive, therefore estimated on observed test results

computed by averaging predictive values over the 10,000 posterior sampled random effects (see section 3.2.2.1). The PA estimates of probabilities to test positive (or negative) according to the sample type and defined infection stages are reported in Table 3.2. In NI fish, the probability to test negative corresponds to DSp. In LI and HI fish, the probability to test positive corresponds to DSe(LI) and DSe(HI), respectively.

3.3.2.1 Non-infected salmon model

In NI fish, significant variation of test results was confirmed at the fish level. The baseline converted PA probability that non-homogenized samples from NI fish test positive (1-DSp) in the reference lab A was 6.4%. Testing in lab C significantly increased (i.e. almost 3.5x) the probability of testing positive (17.8%), whereas testing in lab B slightly decreased this probability (3.3%, not significant). Tissue homogenization increased significantly the probability of testing positive (i.e. almost 3.5x) in lab A (18.0%) and almost doubled in lab C (27.3%), whereas it did not show any effect in lab B (3.3%).

3.3.2.2 Low-infected salmon model

In LI fish, higher variation in test results was revealed at the fish level. The baseline converted PA probability that non-homogenized samples from LI fish test positive (DSe (LI)) in reference lab A was 72.6%. Testing in lab B slightly increased the probability of testing positive (76.6%, not significant). Tissue homogenization

significantly increased the probability of testing positive in lab A (87.2%) and in lab B (90.9%). Regardless of the sample type, all samples tested positive in lab C and the predicted probability of LI fish to test positive was100%.

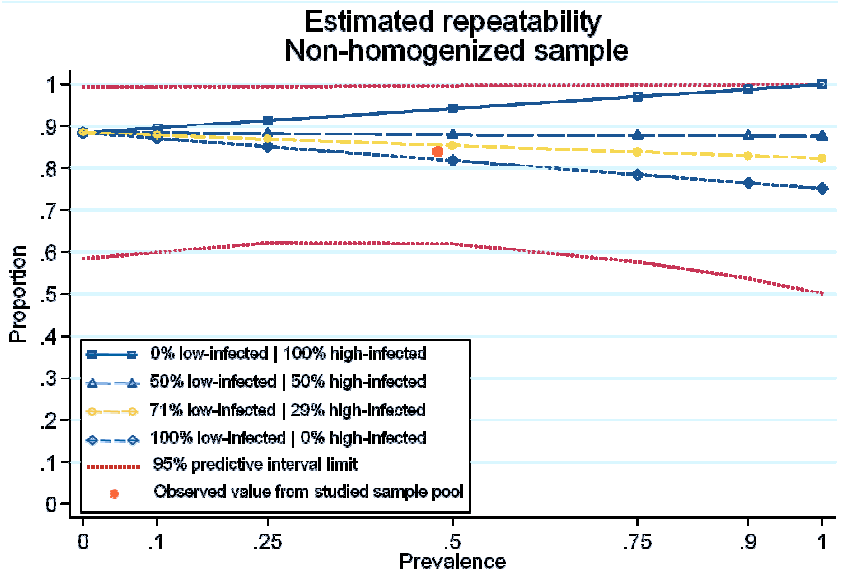
3.3.2.3 High-infected salmon model

Since all 14 HI fish tested positive, regardless of the sample type and processing laboratory, no model was required to predict the probability of HI fish to testing positive; DSe(HI) was assumed to be 100%.

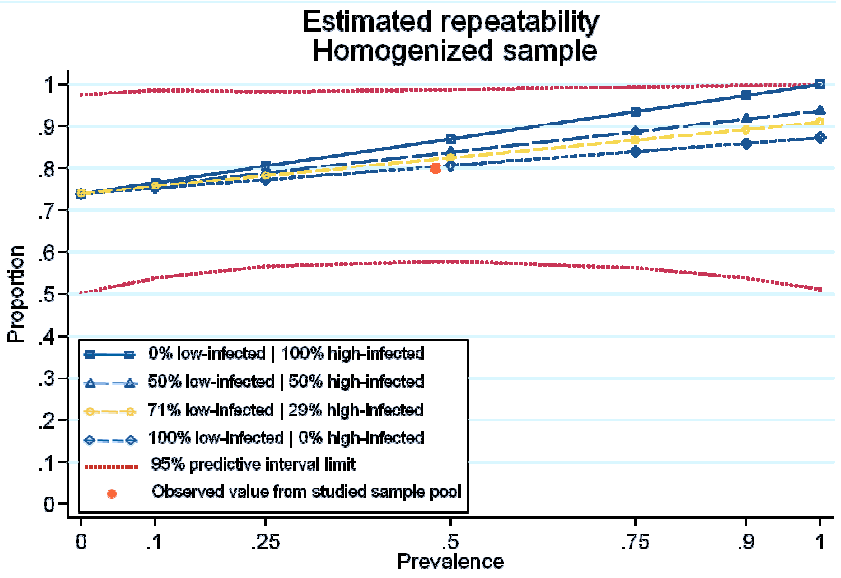
3.3.3 Agreement prediction

3.3.3.1 Estimated agreement

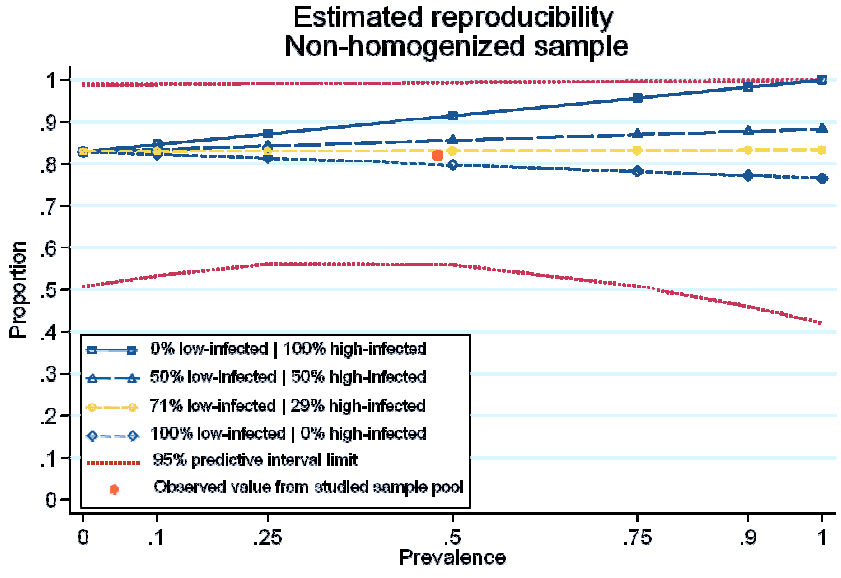
Variation of estimated repeatability (*r*) and reproducibility (*R*) in non-homogenized and homogenized samples across infection prevalences and proportions of LI are illustrated in 4 separate graphs in Fig. 3.3. Based on NI fish (prevalence = 0%) only, estimated *r* and *R* were greater for non-homogenized samples compared to homogenized samples (0.88 and 0.83 vs. 0.74 and 0.73, respectively). For each sample type, estimated *R* was slightly lower than estimated *r*. Increased infection prevalence was associated with an increase in estimated *r* and *R* in homogenized samples, whereas the progression of estimated *r* and *R* in non-homogenized samples depended on the proportion of LI fish in the sample pool. An increased proportion of LI in the group substantially decreased the estimated *r* and, to a lower degree, the estimated *R* for non-



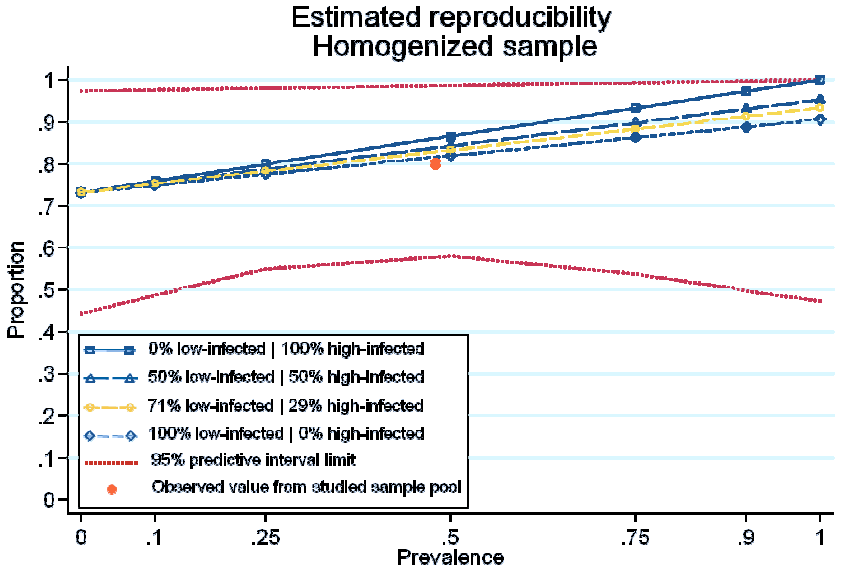
Graph. 3A



Graph. 3B



Graph. 3C



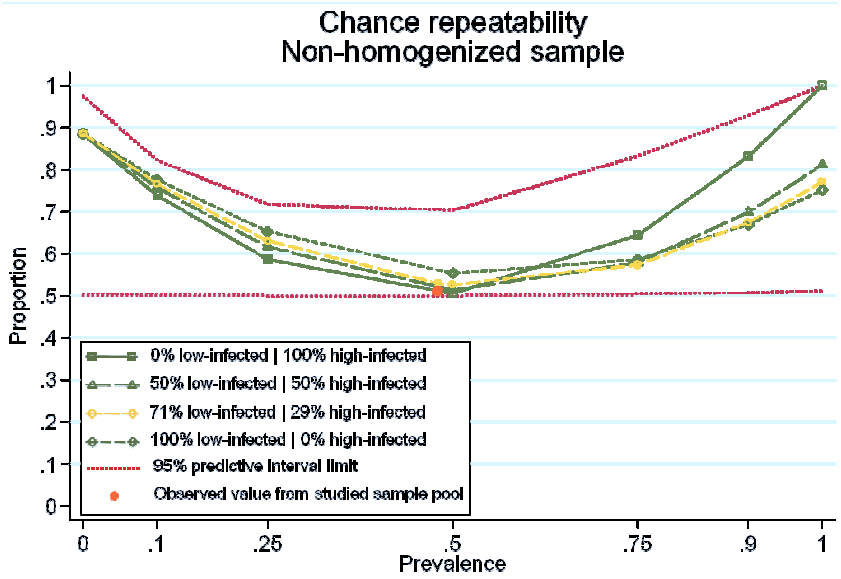
Graph. 3D

Fig. 3.3. Computed estimated agreement of ISAV RT-PCR. Within the reference laboratory for non-homogenized (A) and homogenized sample (B); and reproducibility for non-homogenized (C) and homogenized sample (D) as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circles represent the originally observed estimates under the same testing conditions.

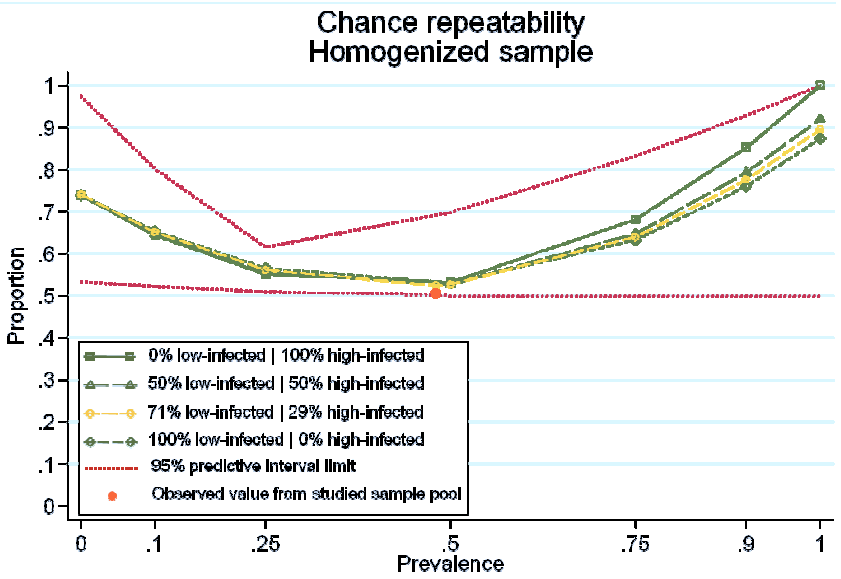
homogenized samples. For instance, estimated *r* and *R* with only LI fish (0.75 and 0.77, respectively) were lower for non-homogenized samples than with NI fish. For homogenized samples, estimated *r* and *R*, based only on LI fish (0.87 and 0.91, respectively), were higher than with NI fish. Estimated *r* and *R* with only HI were perfect (i.e. 100%) regardless of sample type. According to the cone shaped spread of estimated values and associated predictive intervals (Fig. 3.3), estimated agreement in homogenized samples were little influenced by infection stage compared to non-homogenized samples; and predicted *R* seemed to be wider spread than *r*. The initial observed estimates (filled circles) set at the 48% prevalence, according to the PGS, were fairly close to the estimates predicted by the model for 71% of LI among infected fish (Fig. 3.3).

3.3.3.2 Chance agreement

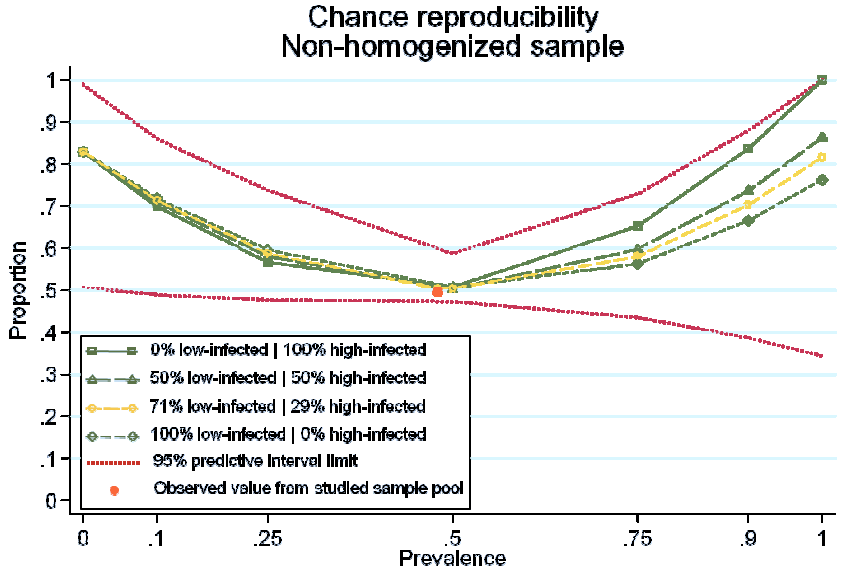
Variation of chance *r* and *R* in non-homogenized and homogenized samples across infection prevalences and proportions of LI are illustrated in 4 separate graphs in Fig. 3.4. All curves showed a convex profile with a minimum prevalence around 50%. Chance agreements were fairly stable across levels of infection when prevalence was lower than 50% and slightly more variable when prevalence was higher. Chance *r* and *R* with only NI fish (prevalence = 0%) were higher for non-homogenized samples compared to homogenized samples (0.87 and 0.81 vs. 0.69 and 0.70, respectively). However, for non-homogenized samples, chance *r* and *R* with only LI fish (0.62 and 0.73, respectively) were lower than with NI fish. For homogenized samples, chance *r* and *R* with only LI fish (0.82 and 0.89, respectively) were much higher than with NI fish.



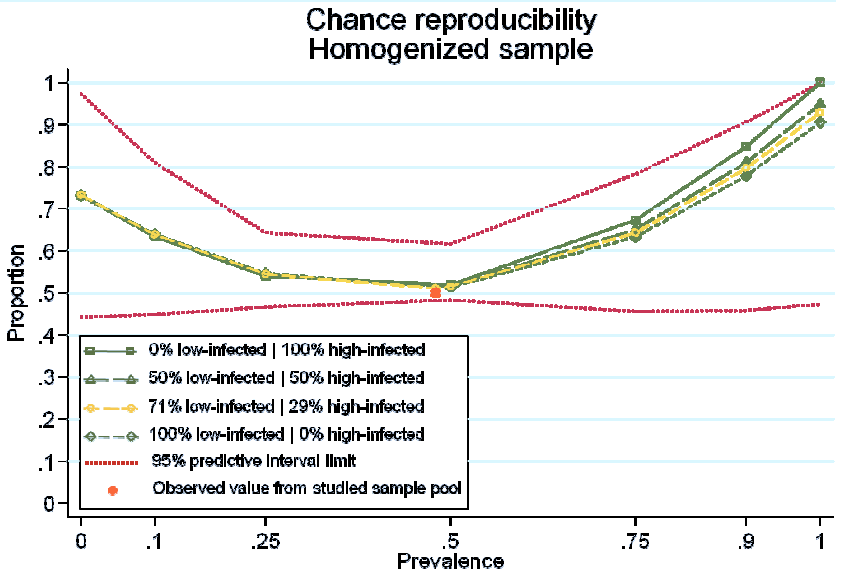
Graph. 4A



Graph. 4B



Graph. 4C



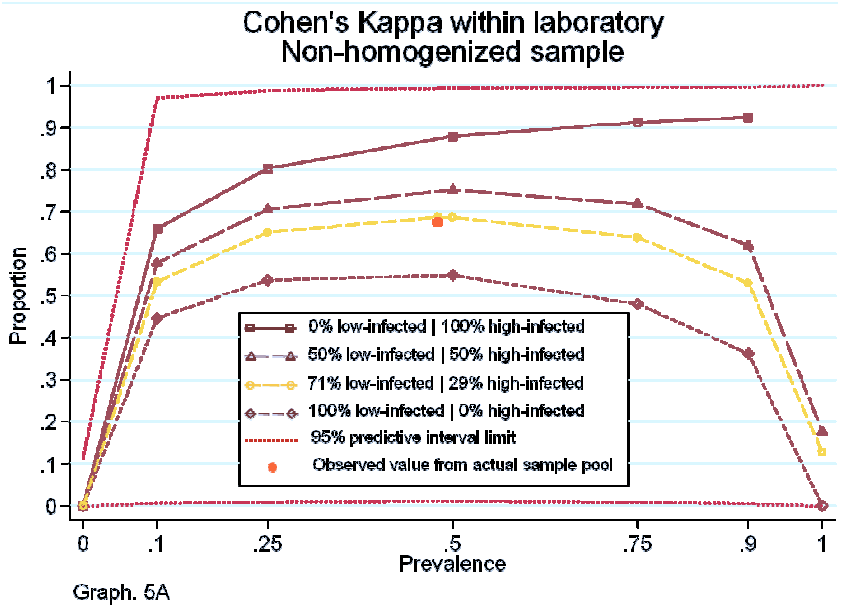
Graph. 4D

Fig. 3.4. Computed chance agreement of ISAV RT-PCR. Within the reference laboratory for non-homogenized (A) and homogenized samples (B); and reproducibility for non-homogenized (C) and homogenized samples (D) as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circles represent the originally observed estimates under the same testing conditions.

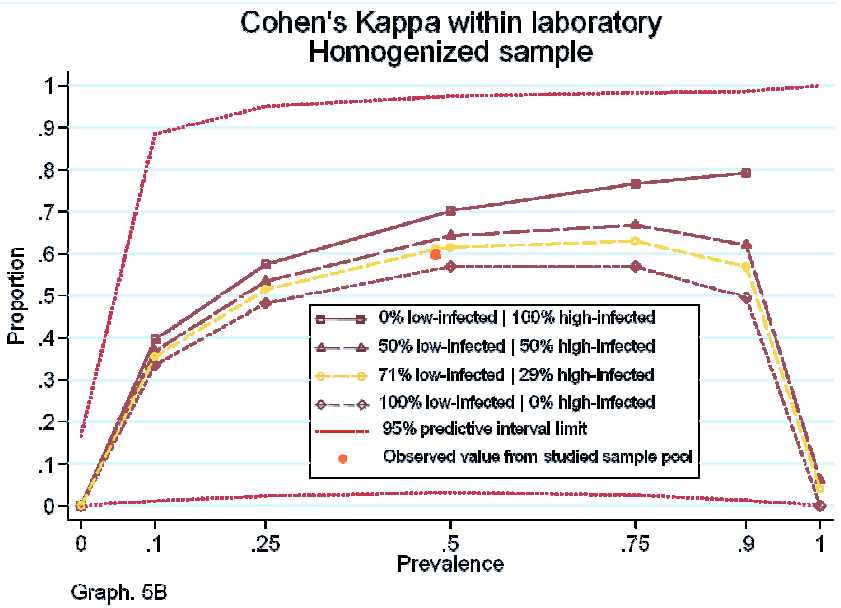
Therefore, profiles for non-homogenized samples looked more symmetrical than for homogenized samples (Fig. 3.4). Chance *r* and *R* with only HI were perfect (i.e. 100%) regardless of sample type. According to the horn-shaped spread of predicted values and associated predictive intervals (Fig. 3.4), chance agreements in homogenized samples were less influenced by infection stage compared to non-homogenized samples, and chance *R* seemed to be also more spread than chance *r*. The initial observed estimates (filled circles) fitted the estimates predicted by the model for 71% of LI among infected fish (Fig. 3.4).

3.3.3.3 Kappa values

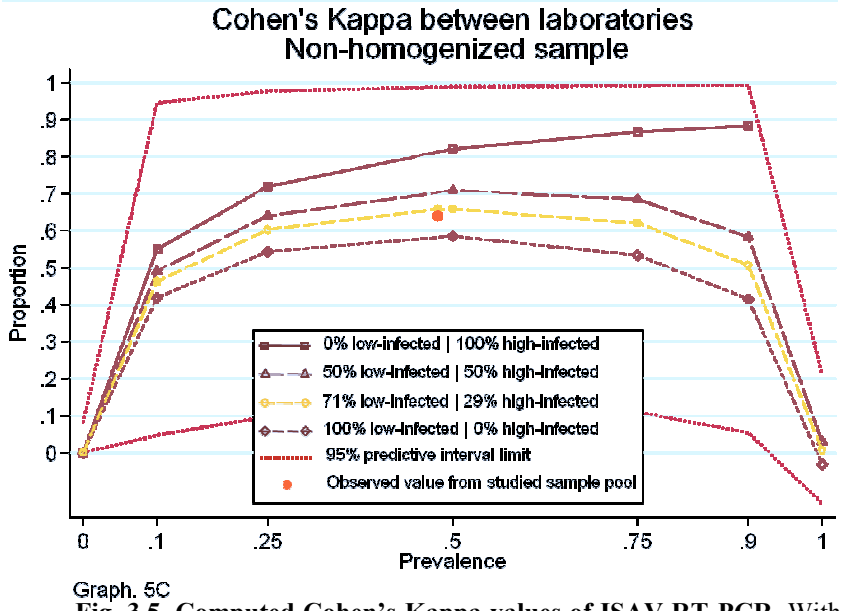
Variation of κ within and among laboratories in non-homogenized and homogenized samples across prevalences and proportions of LI are illustrated in 4 separate graphs in Fig. 3.5. All curves showed a concave profile. Curve profiles were fairly stable when prevalence values ranged from 20 to 80% and κ decreased when prevalence values were extreme. According to the spread of the estimates, κ for homogenized samples was more influenced by prevalence and less influenced by infection stage when compared to non-homogenized samples. The initial observed κ estimates (filled circles) were fairly close to the κ predicted by the model for 71% of LI among all infected fish (Fig. 3.5).



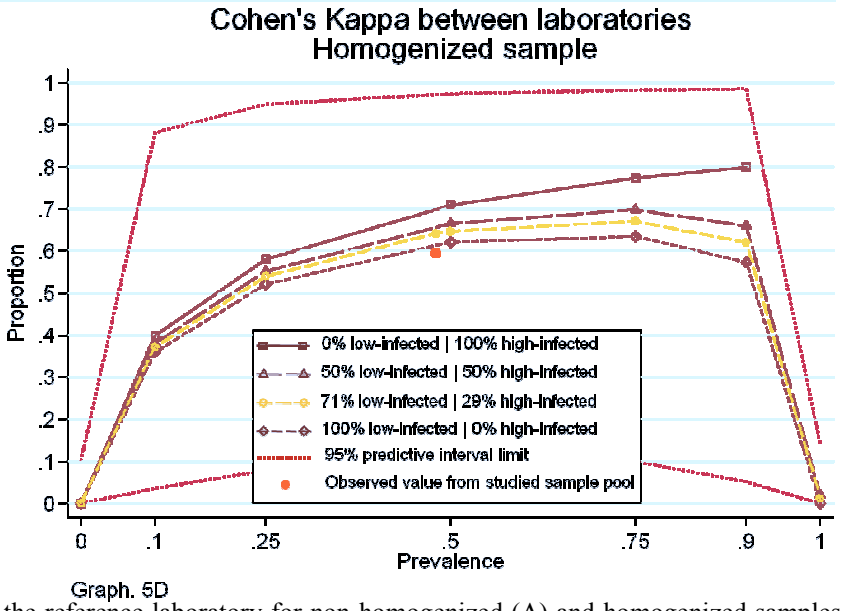
Graph. 5A



Graph. 5B



Graph. 5C



Graph. 5D

Fig. 3.5. Computed Cohen's Kappa values of ISAV RT-PCR. Within the reference laboratory for non-homogenized (A) and homogenized samples (B); and among laboratories for non-homogenized (C) and homogenized samples (D) as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circles represent the originally observed estimates under the same testing conditions.

3.3.4 Alternative estimation to modeling

Direct estimation of probabilities to test positive (or negative), from the dataset split according to the three PGS definitions, are summarized in Table 3.2 for comparison with modelling estimates. Similarly, corresponding agreement estimates using the different approaches (i.e. observed, modelling, and descriptive) are also summarized in Table 3.3. For comparison of prediction, estimated, chance and κ agreement graphs were generated based on descriptive estimates and on PGS, Strict-PGS and Lenient-PGS definition and are presented in the thesis appendices (Appendix 1 to 9, respectively).

3.4 Discussion

3.4.1 Dependence of repeatability and reproducibility

The main objective of this study was to extend the evaluation of the consistency of an ISAV RT-PCR test by predicting r and R across a range of prevalences and infection stage distributions. Adapting Bachmann et al. (2009) approach, we used modelling to predict the probability to yield the same test results from a single farmed Atlantic salmon according to the prevalence of infection in its population of origin, and if infected, the load of virus in its kidney. The modelling approach permitted us to estimate the influence of submission factors on agreement, including the type of sample (i.e. homogenized or not), and the testing laboratory. The developed model identified strong inter-dependence

Table 3.3

Comparison of repeatability and reproducibility estimates of homogenized or non-homogenized samples between originally observed agreement estimates (from Chapter II) and descriptive or modelling population-averaged estimates using the default pseudogold standard classification (PGS) for non-, low- and high-infected fish (NI, LI, HI). Descriptive agreement estimates were added from the two alternative pseudogold classifications (Strict- & Lenient-PSG) for comparison and assessment of the impact of pseudogold definition on prediction.

						Repeatability		Reproducibility	
	Method	Prevalence	%NI	%LI	%HI	Non-homogenized	Homogenate	Non-homogenized	Homogenate
Estimated agreement	Observed					0.840	0.800	0.820	0.800
	Modelling_PGS	0.48	0.52	0.34	0.14	0.856	0.821	0.831	0.829
	Descriptive_PGS	0.48	0.52	0.34	0.14	0.807	0.784	0.807	0.814
	Descriptive _Strict-PGS	0.44	0.56	0.30	0.14	0.858	0.763	0.828	0.774
	Descriptive _Lenient-PGS	0.52	0.48	0.38	0.14	0.805	0.734	0.784	0.742
Chance agreement	Observed					0.494	0.510	0.496	0.500
	Modelling_PGS	0.48	0.52	0.34	0.14	0.529	0.524	0.504	0.511
	Descriptive_PGS	0.48	0.52	0.34	0.14	0.510	0.505	0.496	0.502
	Descriptive _Strict-PGS	0.44	0.56	0.30	0.14	0.510	0.505	0.496	0.502
	Descriptive _Lenient-PGS	0.52	0.48	0.38	0.14	0.510	0.504	0.496	0.503
Cohen's Kappa	Observed					0.674	0.597	0.639	0.595
	Modelling_PGS	0.48	0.52	0.34	0.14	0.687	0.610	0.658	0.642
	Descriptive_PGS	0.48	0.52	0.34	0.14	0.607	0.564	0.617	0.626
	Descriptive _Strict-PGS	0.44	0.56	0.30	0.14	0.709	0.522	0.659	0.545
	Descriptive _Lenient-PGS	0.52	0.48	0.38	0.14	0.603	0.463	0.571	0.481

between agreement and each of these factors.

3.4.1.1Dependence on homogenization

For both NI and LI salmon, homogenisation of the kidney samples significantly increased the probability to test positive compared to the same fish pre-homogenization (Tables 3.1 and 3.2). In NI fish, this led to a higher proportion of false positives and the test DSp moving away from perfection (100%) toward a middle range performance (Table 3.1). Therefore, homogenization decreased estimated and chance agreements in NI fish. The most likely explanation for increased proportion of false positives in homogenates was a sequential contamination from previous homogenized infected samples to subsequent non-infected samples during the homogenization process, as was discussed elsewhere (Chapter II). Regardless of the testing conditions (homogenization and laboratory), Cohen's kappa (κ) values were nil for NI because agreement was only due to chance (κ measures agreement beyond chance; Dohoo et al., 2009).

The increased probability to test positive in LI fish reflects a higher proportion of true positive (i.e. DSe(LI) more extreme and closer to 100%) and therefore an increase in estimated and chance agreements. Again, κ for LI were close to nil since agreement was by chance alone. The higher probability to test positive in homogenates from LI salmon might be explained by a better yield of viral RNA during the extraction facilitated by a more efficient cell disruption due to the previous homogenization. Further investigations should be conducted at the bench level to evaluate if homogenization would improve the analytical sensitivity of this assay.

All agreement estimates for homogenized samples revealed a much lower variation across different proportions of LI. The apparent improvement of robustness by homogenization in LI fish is consistent with the suspected heterogeneous distribution of viral particles during early infection stages. Previously discussed in Chapter II, this suggests a multifocal distribution of infectious clusters of low numbers of viral particles in the kidney of LI salmon. Higher consistency of virus detection in homogenates may also be due to over contamination of samples. However, this aspect could not be assessed in this study.

The decreased DSp in homogenates would support the use of homogenization toward a confirmatory diagnostic test and away from a screening use. This may, however, be impractical, and homogenization showed encouraging consistency for LI fish which are targeted by surveillance programs. Homogenization of tissue samples has diverse relevant applications for ISAV control programs (e.g. pooling of samples, production of aliquoted reference and control materials, laboratory proficiency testing). Further development and standardization of the homogenization process is needed to optimize the assay's performance and consistency.

3.4.1.2 Dependence on processing laboratory

In this study, *R* was not reported for separate pair-wise comparisons between two laboratories, but as an average of the three pair-wise estimates among the three participating laboratories (Labs A, B and C). Pooled *R* showed less variation than particular pair-wise estimates which would be expected to cover wider predictive

intervals. Overall, the average *R* showed less influence from proportion of LI fish than *r*. As an average value, this parameter is less susceptible to variation, but wider predictive intervals confirmed that it could take more extreme values. This observation is consistent with the concept that *R* carries additional degrees of variability compared to *r* due to laboratory divergences, even when assessed as an average.

Lab C had a significantly higher probability to test positive than the two other laboratories, based on samples from NI fish (i.e. lower DSp) (Tables 3.1 and 3.2). This reduced performance in NI fish might be explained by increased cross-contamination (Wilson, 1997) in this particular laboratory and might result in a lower *r* in this particular category of salmon (not assessed here). Conversely, the probability to test positive in lab C for samples from LI fish was perfect (DSe(LI) = 100%) which might increase the overall *r* (not assessed here). Since the agreement is perfect for infected fish in lab C (DSe(LI) and DSe(HI) = 100%), the overall *r* in this particular laboratory would depend only on the proportion of NI of fish in the tested pool (1 - prevalence). As suggested by Bland and Altman (1986), the lack of agreement of lab C with the two other laboratories may be explained by poor *r* in lab C. However, considering results from both types of samples (homogenized and non-homogenized), the observed agreement within lab C was acceptable (0.85 in Chapter II). Therefore, the poor observed agreement between this facility and the others laboratories described in Chapter II might be explained by consistent divergent performances in lab C. Albeit agreement between lab A and B was more acceptable (Chapter II), clearly the method was difficult to transfer between laboratories without significantly changing test performances. Reasons for poor

transferability of this method, including factors associated with laboratory organization and staff experience, were discussed in detail elsewhere (Chapter II).

3.4.1.3 Dependence on infection prevalence

R and *r* within each infection category (infected and non-infected) are intrinsically associated with test performances (DSe and DSp, respectively). If agreement differs across infection categories, the overall agreement would result in a prevalence-weighted average. The establishment of a PGS allowed us to estimate category-specific agreement, and confirmed the discrepancy of agreement among infection stages (Table 3.2). Similar to overall accuracy or efficiency (Alberg et al., 2004), it was expected that test agreement strongly varies when (i) DSe and DSp differ much from each other and/or (ii) the prevalence of the targeted population deviates from 50%. By category weighting, agreement was predicted across prevalences and revealed substantive levels of variation according to the submission factors. Application of these predicted agreements depend on the specific utilization of the assay. If the purpose of the test is routine surveillance, the infection prevalence in the targeted population is likely to be low, and predicted precision would provide better fit to true agreement values. Alternatively, if the intended purpose of the test is diagnostic confirmation, the infection prevalence is likely to be high and the predicted estimates would strongly change the expectation of test agreement based on the proportion of infection stages.

Across prevalences, estimated agreement showed monotonic linear profiles (Fig. 3.3), whereas chance agreement and κ values revealed convex and concave profiles,

respectively (Fig. 3.4 and 3.5). Chance agreements, regardless of the testing conditions, reached a minimum around 50% prevalence which was close to the assumed prevalence of the studied pool of samples (according to the PGS). Although initial observed agreements appeared low (Chapter II), the interpretation should consider that the actual agreement evaluation was evaluated under the most challenging conditions. Representing the test result agreement due to chance alone (Dohoo et al., 2009), chance agreement of an imperfect test (DSe and DSp < 100%) is the lowest when the test has the highest number of misclassifications. For instance, when DSp and DSe are equal, chance agreement is minimum (.50) at exactly 50% prevalence. In this study, however, Dsp and DSe in LI and HI fish differed and minima were reached at prevalences moving away from 50%. Computation of κ from Eq. (10) was mainly influenced by the difference between estimated and chance agreement (numerator). In theory, κ reaches a maximum approximately when chance agreement is minimal under the assumption that estimated agreement is fairly stable (Fig. 3.3, 3.4 and 3.5). The resulting concave profile of κ across prevalences was consistent with previous descriptions of κ variations (Guggenmoos-Holzmann, 1996). This is another example of the fair stability of κ in mid-prevalence ranges and the strong, almost symmetrical, decrease of κ for extreme prevalences (i.e. below 20% and above 80%). In these extreme prevalence ranges, interpretation of κ is difficult which supports the recommendation to use samples from population close to 50% prevalence to evaluate *r* and *R*.

The predictive approach for agreement across different prevalences allowed us to extrapolate test consistency to external populations. It assumes, however, that the agreement is constant within each infection category. Advanced explorations of test

performances tend to indicate that DSp and DSe are not constant parameters and may vary according to population factors such as infection prevalence (Greiner & Gardner, 2000). The impact of particular biological factors on test performances, such as proportions of infection stages, were further investigated to assess the possibility that agreement could fluctuate within infected fish for a similar prevalence.

3.4.1.4 Dependence on proportion of infection stages

The probabilities to test positive (or negative) in the three infection categories differed substantively (Table 3.2), and resulted in agreement discrepancies among the three stages. The perfect test agreement with HI confirms the good performances of RT-PCR in salmon with high concentrations of ISAV. Encouraging for advanced outbreaks, the intended purpose of this assay, however, was also to target detection of early ISAV stages during surveillance. Therefore, prediction of overall agreement for surveillance has to be adjusted for the proportion of LI fish (assuming that agreement is constant within each infection stage).

DSe is closely associated with the detection limit of the assay (i.e. analytical sensitivity), and therefore the probability of detection in LI fish was expected to be lower (low number of viral particles) or less consistent than in HI fish (Fig. 3.1). Indeed, probabilities to test positive in LI fish (DSe(LI)), and therefore agreements, were far from perfect except in Lab C (Table 3.2). This further supports the hypothetical heterogeneous distribution of viral particles (clustered) in early infection stages of ISA. With non-homogenized tissue, the operator can also expect lower agreement in LI than in NI

specimen (Fig. 3.3). According to Fig. 3.5, similarly to NI, agreement in LI is only due to chance ($\kappa = 0$), regardless of the sample type. Therefore, the RT-PCR performances decrease with viral concentration in non-homogenized samples. Most binary diagnostic tests are based on the dichotomization of individuals according to a continuous underlying trait (i.e. target concentration) (Brenner and Gefeller, 1997). It is then recommended that test performance (DSe) and agreement be preliminarily investigated (at the bench level) across a gradient of analyte concentrations. For instance, a regression equation associating agreement and target concentration may be implemented to predict agreement within infected individuals.

Ultimately, using agreement estimates specific to infection stage (assumed constant within stages) and field evidence of the association between prevalence and relative proportions of infection stages, agreement can be predicted for different configurations to validate the agreement estimation and allow extrapolation to new populations.

3.4.2 Validity of the modelling approach

The implementation of a modelling approach to predict agreement revealed very practical applications but also required several assumptions. Briefly, the defined PGS was assumed to correctly classify fish into the three infection stages; and stage-specific agreement estimates were assumed to be constant across populations and also across infection prevalences.

3.4.2.1 Comparison with observed and descriptive estimates

Following confirmation of MCMC chains convergence (see section 3.2.2.2), indication of goodness-of-fit was assessed visually by the proximity of predicted values with originally observed values in each graph (Fig. 3.3, 3.4 and 3.5). Overall, modelling tended to slightly overestimate agreement, though these values were still very close to each other. Compared to the originally observed estimation of agreement (Chapter II), modelling estimation enabled assessment of dependence among observations (e.g. from the same fish) and assessment of submission factors (e.g. homogenization) on test agreement.

Tables 3.2 and 3.3 also compared probabilities and agreement obtained from different methods, including observed, modelling and descriptive estimates. Descriptive estimation is an alternative approach to evaluate probabilities to test positive (or negative) directly from the fish results classified according to PGS definition. This descriptive approach, however,neglected the dependence of duplicated test results coming from the same fish. Estimated probabilities were very similar between modelling and descriptive approaches (Table 3.2), and the corresponding agreement estimates stayed close to initially observed values, although they could differ slightly from each other (Table 3.3). The alternative estimation approach (descriptive) does not provide predictive intervals for agreement, thus restricting the assessment of the full range of future values. The predicted intervals indicated in Fig. 3.3-3.5 were very wide, in part because of the substantial between-fish random variation, but also because they corresponded to 95% predictive ranges for a sample size of one salmon. If agreement was computed from a

larger sample, its predictive bounds would become narrower. Basically, the modelling approach requires a modest extra analytical effort but provides valuable additional information.

3.4.2.2 Sensitivity of prediction to the pseudogold standard definition

Both predictive approaches (i.e. modelling and descriptive) rely on the definition of the PGS to classify fish into infection categories whereby the accuracy of the classification may affect the validity of the prediction. For example, the test DSe seemed to be lower in lab C than in other laboratories based on the PGS classification of fish, or alternatively the assay analytical sensitivity may be greater in this laboratory, allowing detection of very low virus levels. Originally, PGS was set that any observation had equal weight and the fish classification was based on evidence supported in the data. To be classified as infected, a salmon had to be positive in at least two laboratories (excluding lab A duplicates), and in the two sample types (non-homogenized and homogenized) (i.e. at least four positive results out of the six conventional RT-PCR runs) (Fig. 3.1). The subsequent classification differentiated low- and high-infected fish based on real-time RT-PCR results using a Ct cutpoint set at the mid-point of the expected range for Ct values. We acknowledge that this value is arbitrary and requires further investigation to explore ISAV levels existing in salmon at different stages of infection and correlate them to Ct values. To explore any major influence of PGS definition on agreement prediction, we compared changes in probabilities and agreement estimates using the descriptive approach for probability estimation (see section 3.2.5) based on alternative PGS

definitions (Tables 3.2 and 3.3). Compared to PGS, Strict-PGS classified less fish as infected which resulted in increased DSp and decreased DSe, while Lenient-PGS classified more fish as infected and resulted in decreased DSp and increased DSe (Table 3.2). Graphical prediction of agreement for both PGS alternative definitions revealed approximately the same trends compared to the initial PGS (data not shown), while comparison with initially observed estimates revealed small differences (slightly larger than for PGS) (Table 3.3). Albeit the PGS criteria may change some salmon infection status, the overall agreement trends did not change substantially when the PGS definition was changed.

3.4.2.3 Dependence of diagnostic sensitivity and specificity on infection prevalence

Prediction of agreement assumed that agreement within each infection category was constant and that test performances did not depend on the factors varying across populations. Although, it is commonly accepted that DSp (or DSe) are constant parameters within each infection category and do not vary with population factors such as prevalence, various studies have challenged this postulate (e.g. Leeflang et al., 2009) and the invalidity of these assumptions may introduce estimation bias for predicted agreement. Further discussions on this issue and methods to adjust the estimation procedure are proposed in Appendix 15.

3.5 Conclusion

Conventional descriptions of *r* and *R* have limited extrapolation, in particular when test agreement is expected to vary with the stage of infection and therefore to be dependent on prevalence. External validity of *r* and *R* estimation requires the possibility to predict the change in agreement across prevalences and infection stage distributions. Utilization of multilevel modelling procedures in a Bayesian framework allowed for assessment of the effect on agreement of submission factors as well as accounting for dependence among samples from the same fish. Briefly:

- Homogenization of salmon tissue samples tended to improve the detection and agreement in early infected salmon with ISAV, which supports a hypothetical heterogeneous distribution of ISAV in an infected kidney. Higher proportions of positives with homogenates, however, might also result from cross-contamination in non-infected fish.

- ISAV RT-PCR in salmon was not always transferred with success (e.g. lab C) and may require further protocol standardization.

- *r* and *R* increased in a linear fashion with the proportion of highly infected salmon and decreased with the proportion of non-homogenized lowly infected samples for ISAV. RT-PCR agreement was perfect with samples from highly infected salmon.

- Kappa profiles revealed concave patterns, with kappa maxima reached approximately at 50% prevalence of ISAV infection in salmon, and sharp drops for extremes in prevalence (prevalence < 20% or > 80%). Agreement for non-infected (0% prevalence) and lowly infected fish was mainly due to chance.

3.6 References

Alberg, A.J., Park, J.W., Hager, B.W., Brock, M.V., Diener-West, M., 2004. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J. Gen. Intern. Med. 19, 460-465.

Anonymous, 2000. ISA hits the Faroes. Fish Farming International. 27, 47.

Bachmann, L.M., ter Riet, G., Weber, W.E., Kessels, A.G., 2009. Multivariable adjustments counteract spectrum and test review bias in accuracy studies. J. Clin. Epidemiol. 62,357-361.

Begg, C.B., 1987. Biases in the assessment of diagnostic tests. Stat. Med. 6, 411-423.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1, 307-310.

Björk, J., Grubb, A., Nyman, U., (2009). Variability in diagnostic accuracy can be estimated using simple population weighting. J. Clin. Epidemiol. 62, 54-7.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C.W., 2003. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin. Radiol. 58, 575-580.

Bouchard, D.A., Brockway, K., Giray, C., Keleher, W., Merrill, P.L., 2001. First report of infectious salmon anaemia (ISA) in the United States. Bull. Eur. Assoc. Fish. Pathol. 21, 86-88.

Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat. Med. 16, 981-991.

Brooks, S.P., Gelman, A., 1998. Alternative methods for monitoring convergence of iterative simulations. J. Comp. Graph. Stat. 7, 434-455.

Cleophas, T.J., Droogendijk, J., van Ouwerkerk, B.M., 2008. Validating diagnostic tests, correct and incorrect methods, new developments. Curr. Clin. Pharmacol. 3, 70-76.

Dohoo, I., Martin, W., Stryhn, H. (Eds.), 2009. Veterinary Epidemiologic Research. 2nd ed., AVC Inc., Charlottetown, Canada.

Gelman, A., 1996. Inference and monitoring convergence, Chapter 8. In: Gilks, W., Richardson, S., Spiegelhalter, D. (Eds.), Markov Chain Monte Carlo in practice. Chapman et Hall, London, UK, pp. 131-140.

Godoy, M.G., Aedo, A., Kibenge, M.J., Groman, D.B., Yason, C.V., Grothusen, H., Lisperguer, A., Calbucura, M., Avendaño, F., Imilán, M., Jarpa, M., Kibenge, F.S., 2008. First detection, isolation and molecular characterization of infectious salmon anaemia virus associated with clinical disease in farmed Atlantic salmon (*Salmo salar*) in Chile. BMC Vet. Res. 4, 28.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 42, 2-22.

Guggenmoos-Holzmann, I., 1996. The meaning of Kappa: concepts of reliability and validity revisited. J. Clin. Epidemiol. 49, 775-782.

ISO International Standard 5725-1, 1994. Accuracy (trueness and precision) of measurement methods and results. Part 1: General principles and definition. International Organisation for Standardisation

(ISO), ISO Central Secretariat, 1 rue de Varembé, Case Postale 56, CH - 1211, Geneva 20, Switzerland.

Leeflang, M.M., Bossuyt, P.M., Irwig, L., 2009. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J. Clin. Epidemiol. 62, 5-12.

McClure, C.A., Hammell, K.L., Stryhn, H., Dohoo, I.R., Hawkins, L.J., 2005. Application of surveillance data in evaluation of diagnostic tests for infectious salmon anemia. Dis. Aquat. Organ. 63, 119-27.

Mullins, J.E., Groman, D., Wadowska, D., 1998. Infectious salmon anaemia in salt water Atlantic salmon (*Salmo salar* L) in New Brunswick, Canada. Bull. Eur. Assoc. Fish. Pathol. 18, 110-114.

Nérette, P., Dohoo, I., Hammell, L., Gagné, N., Barbash, P., MacLean, S., Yason, C., 2005. Estimation of the repeatability and reproducibility of three tests for infectious salmon anaemia virus. J. Fish. Dis. 28, 101-110.

Office International des Epizooties, 2008. OIE Quality Standard and Guidelines for Veterinary Laboratories: Infectious Diseases. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 70pp.

Office International des Epizooties, 2009a. OIE Aquatic Animal Health Code. 12th Edition. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 99-104.

Office International des Epizooties, 2009b. Manual of Diagnostic Tests for Aquatic Animals 2009. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 10-30.

Ransohoff, D.F., Feinstein, A.R., 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N. Engl. J. Med. 17, 926-930.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009) *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Rodger, H.D., Turnbull, T., Muir, F., Millar, S., Richards, R., 1998. Infectious salmon anaemia (ISA) in United Kingdom. Bull. Eur. Assoc. Fish. Pathol. 18, 115-116.

Smyth, G., 2005. Numerical Integration. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of statistics. 2nd ed., John Wiley & Sons, Inc.

Spiegelhalter, D.J., Thomas, A., Best, N.G., 2003. WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit.

Thorud, K.E., Djupvik, H.O., 1988. Infectious salmon anaemia in Atlantic salmon (*Salmo salar* L). Bull. Eur. Assoc. Fish. Pathol. 8, 109-111.

Yang, I., Becker, M.P., 1997. Latent variable modeling of diagnostic accuracy. Biometrics. 53, 948-958.

Van den Bruel, A., Cleemput, I., Aertgeerts, B., Ramaekers, D., Buntinx, F., 2007. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J. Clin. Epidemiol. 60,1116-1122.

Wilson, I.G., 1997. Inhibition and facilitation of nucleic acid amplification. Appl. Environ. Microbiol. 63, 3741-3751.

Zeger, S.L., Liang, K.Y., Albert, P.S., 1988. Models for Longitudinal Data: A Generalized Estimating Equation Approach. Biometrics 44, 1049-60.

Chapter IV: USE OF A THIRD CLASS IN LATENT CLASS MODELLING FOR DIAGNOSTIC TEST EVALUATION: APPLICATION TO THE EVALUATION OF FIVE INFECTIOUS SALMON ANAEMIA VIRUS DETECTION ASSAYS

Abstract

More than two classes (diseased and non-diseased) have been previously used to address the assumption of conditional test independence when using latent class model (LCM) to evaluate diagnostic test performance in the absence of a gold standard. The addition of supplementary class was investigated in this study to also address the assumption of constant diagnostic sensitivity (DSe) and specificity (DSp) across populations. Applied to 5 detection methods for infectious salmon anaemia virus (ISAV) using 400 Atlantic salmon sampled from 4 populations, a three class LCM was implemented in a Bayesian framework with further adjustment for test dependence. The model successfully recognized a third class of fish with substantially differing test performances. Only the nucleic-acid amplification assay could detect this additional class of fish. The definitive identity of the third class (infected or not) was subject to discussion and required further knowledge about the infection dynamics at the fish level. The obtained estimates of test DSe and DSp only apply to populations of Atlantic salmon farmed in Atlantic Canada. Regardless of the potential applications of multiple class estimates, selection of the appropriate assay or testing strategy for ISAV infection requires further detailed information of the infection dynamics at the population level.

4.1 Introduction

4.1.1 Latent Class Modelling for diagnostic evaluation and underlying assumptions

Conventionally, the evaluation of diagnostic test performances has involved a comparison of a candidate test with perfect (gold standard) or imperfect reference methods (Dohoo et al., 2009). In the absence of a reference standard, latent class models (LCM) have become the analytical tool of choice. Pioneers in the application of LCM to diagnostic evaluation, Hui & Walter (1980) presented key assumptions for using LCM: (i) at least two dichotomous assays should be applied to the same individuals from at least two populations (with different assumed prevalences), (ii) the tests must be conditionally independent given the infection/disease status; and (iii) the operating characteristics of the tests must be constant across populations. In a Bayesian framework, the requirement for at least 2 populations can be relaxed using a single population when prior information about the prevalence and test properties are available (Dendukuri & Joseph, 2001). However, the prior of a specified parameter may have more impact on the posterior distribution of some parameters than the data itself (Neath & Samaniego, 1997). As an alternative, it was recommended to increase the number of tests and/or the number of populations or to stratify the one(s) available (Toft et al., 2005). For instance, a minimum of 4 tests is necessary for a single sampled population (Dendukuri & Joseph, 2001).

4.1.2 Validity of the test independence conditional on the infection/disease status

Conditional independence between two tests means that the probability of one test result is not influenced by the result of the other test (Dohoo et al, 2009). Tests that use similar techniques and/or measure the same biological trait are expected to be conditionally dependent (Gardner et al., 2000). The assumption of test conditional independence may not be reasonable (Brenner, 1996; Hui & Zhou, 1998) and may result in strongly biased estimations (Vacek, 1985; Torrance-Rynard & Walter, 1997). Several methods have been developed to account for conditional dependence (Hui & Zhou, 1998; Toft et al., 2005; Branscum et al., 2005).

The first method was indirectly developed by Vacek (1985) and it involves adding covariance terms to the model (Dendukuri & Joseph, 2001; Georgiadis et al., 2003). With this approach, one test can depend on two other tests only if these other tests do not depend on each other (Nérette et al., 2008a). Later, utilization of log-linear models was described by including first degree interaction terms between pairs of tests in the model (Espeland & Handelman, 1989; Yang & Becker, 1997). Although interaction effects can be included as fixed or random, this assumes that at least one of the interactions is significant and the interpretation of the final model becomes complex (Hui & Zhou, 1998; Toft et al., 2005). Another approach was described by Qu et al. (1996) by extending the LCM with random effects to account for correlation among tests. A major assumption of mixed models is that random effects representing dependence between tests follow the same normal distribution, which might not be reasonable (Toft et al.,

2005). Alternatively, Rindskopf & Rindskopf (1986) suggested relaxing the assumption of test conditional dependence by including additional classes in the model.

Binary tests may be assumed to dichotomize the measure of an underlying continuous trait of the degree of infection/disease (Brenner & Gefeller, 1997). However, This continuum, however, may be differentiated into three or more categories (Rindskopf & Rindskopf, 1986). The inclusion of more than two classes in LCM allows for different levels of infection/disease that correspond to different observed testing patterns giving a better fit of the data (Formann, 1994). Within each class, the assumption of conditional independence may therefore hold. In general, the addition of terms or classes increases the number of model parameters which might compromise the identifiability of the LCM.

4.1.3 Validity of the assumption of constant classification across populations

Often incorrectly referred to as accuracy, the trueness of an assay is defined as the degree of agreement between the average value obtained from a large series of test results and an accepted reference value (ISO 5725-1, 1994). For binary outcomes, this is expressed by the test overall efficiency (Ef) computed as the proportion of specimens that are truly classified. Ef is a prevalence-weighted average of the proportion of true positive in infected/diseased (D+) individuals and the proportion of true negative in non-infected/non-diseased (D-) individuals (Alberg et al., 2004). Classification performances of a test are, therefore, preferentially reported separately for D+ and D- individuals. Conventionally, the diagnostic sensitivity (DSe) refers to the proportion of positive test results among D+ individuals (true positive fraction). Similarly, the diagnostic specificity

(DSp) refers to the proportion of negative test results given the individuals are D- (true negative fraction) (Yerushalmy, 1947).

Contrary to common understanding, DSe and DSp may not be constant parameters and may vary across populations. Greiner & Gardner (2000) considered DSe and DSp as population parameters that vary within and between populations according to the distribution of covariate factors that influence the biology of the infection/disease. For instance, stage and severity of disease are factors commonly suspected to influence DSe. In pathologic screening, DSe is expected to be greater for large lesions than for small lesions (Begg, 1987). Also, compared to early stages, the proportion of animals in advanced stages of an infectious disease testing positive is expected to be higher in high prevalence populations (Greiner & Gardner, 2000), and DSe is expected to increase with prevalence. Conversely, DSp is suspected to be highly impacted by the probability of cross-contamination (Wilson, 1997). For instance, high prevalence populations are expected to generate more opportunities for cross-contamination. Therefore, DSp is expected to decrease with prevalence. In theory, it would be recommended to estimate DSe and DSp specifically for a targeted population using samples randomly collected (i.e. representative of biological covariates) (Johnson et al., 2009). Nevertheless, the disease pattern expressed in a population changes over time. In addition, to be identifiable, LCM may require more than 2 populations with different prevalences (no prior information available). As the mixture distribution of infection/disease stages varies across populations, LCM may result in a biased pooled estimation of DSe (Toft et al., 2005). In a 2-test/2-population scenario, the LCM estimate of DSe is an average of the two population-specific DSe only when one test has perfect DSp (Johnson et al., 2009).

One alternative approach is to increase the number of classes and assume constant DSe (or DSp) within each class, as previously suggested by Rindskopf & Rindskopf (1986) for conditional independence. The proportion of infected animals in each class would vary across the studied populations, while the basic assumptions of constant DSe and DSp are maintained within each sub-class.

4.1.4 Application of LCM to infectious salmon anaemia virus detection

Infectious Salmon Anaemia virus (ISAV) is an Orthomyxovirus, genus Isavirus, causing high mortality in Atlantic salmon, *Salmo salar* L., and except in a few salmon production areas (e.g. British Columbia, Canada), represents a serious threat to the economic sustainability of aquaculture industries around the world (Godoy et al. 2008). ISAV is listed as a reportable disease by the World Organisation for Animal Health (OIE, 2009a) and by the Canadian National Aquatic Animal Health Program which requires that ISAV assays be evaluated and validated. According to the biology of the disease, it is expected that DSe of ISAV tests varies across different prevalence level populations since prevalence of ISAV has been associated with the severity of infection (i.e. proportion of mortalities increased faster than infection prevalence) (Gustafson, 2005). Among the numerous evaluations for ISAV diagnostic tests, three studies have been conducted using LCM (Nérette et al., 2005; Gustafson et al., 2008; Nérette et al., 2008a). Discussing the validity of Hui & Walter assumptions, Nérette et al. (2005) observed variation in test performances with a decrease of DSp when ISA prevalence increased. This inconsistency in classification was further analyzed and confirmed, and some degree of conditional dependence among test pairs was additionally detected (Nérette et al.,

2008a). Although conditional dependence was accounted for in the final evaluation, the variation of DSp across populations was not addressed (Nérette et al., 2008a). Furthermore, no investigation of the model assumptions was reported in Gustafson et al. (2008).

4.1.5 Objectives

The objectives of this study are twofold. First, this study was conducted to evaluate the test characteristics of a recently designed conventional reverse-transcriptase polymerase chain reaction (RT-PCR) assay for ISAV in Atlantic salmon. As recommended in OIE guidelines (OIE, 2009b), the aim of this evaluation was to determine the test's fitness for a specific purpose. The components of the "*purpose-status-test"* triangle, as outlined in (Mintiens et al., submitted), were clearly identified and defined before commencing the evaluation. The intended *purpose* of the test was to demonstrate freedom from ISAV (surveillance) in a population (cage, site, bay, region, Canada). The targeted *status* for detection (diagnostic target) was an Atlantic salmon, *Salmo salar* L., infected with any ISAV genotype regardless of disease. Here the concept of infection is ambiguous and includes carrier states since active viral particles can be detected in a susceptible host before and after the infection *per se*. Although the test of interest was the RT-PCR assay, four additional detection methods, including real-time or quantitative RT-PCR (qRT-PCR), virus isolation (VI), indirect fluorescent antibody test (IFAT), and lateral flow immunoassay (LFI), were used in parallel on the same fish sampled from 4 separate populations to ensure the LCM identifiability.

The secondary objective of this study was to build a LCM for tests for ISAV in salmon that included a third class in a flexible Bayesian framework to explore and account for expected variation in classification performance across populations and address conditional dependence of misclassification error. Adapting the multiple latent variable model recently described by Dendukuri et al. (2009), we further measured the conditional dependence among tests by including covariance terms in the model. The variation of RT-PCR DSp across populations observed by Nérette et al. (2008) could be explained by a differing diagnostic objective where only fish carrying virulent ISAV were targeted. Explained by a different analytical specificity, RT-PCR can detect RNA of avirulent ISAV particles (i.e. HPR0) when other tests may not. Hence, when a sample tested positive only to RT-PCR in Nérette et al. (2008) study, the fish was considered non-infected by the model when it may be truly infected with avirulent particles. Depending on the proportion of this type of samples, the DSp of RT-PCR varied across populations. In this study, the diagnostic target was ISAV infection including any viral genotype (including HPR0) and instead we expected DSe to vary across populations. Therefore, we first assumed that the added class was a supplementary class of infected salmon.

The justification of a third latent class from patterns in the data as well as its possible interpretations are investigated and discussed in detail. This reflects our view that multiple latent class models may seem attractive for modelling but need to be well-founded in the data in order to obtain meaningful results. For instance, the relative presence of the third class in the data would increase its ability to be detected. Thus we believe our study could serve as a case study for further applications of the methodology.

4.2. Material and methods

4.2.1 Target and study populations

The target population for this study was farmed Atlantic salmon, *Salmo salar* L., grown in sea cages in the province of New Brunswick, Canada. Other farmed salmon exposed to similar environmental and viral genotype conditions, such as those in Nova Scotia and Newfoundland, are considered a first degree external population. Wild Atlantic salmon might be included as a second degree external population, assuming external validity.

The studied populations consisted of farmed Atlantic salmon of 2004 and 2005 year class, grown in polar circle cages in the Bay of Fundy, New Brunswick, Canada. Four populations were identified with different infection prevalence and representing a range of infection stages according to clinical and historical information (McClure et al., 2004): (i) a near-zero prevalence population included apparently healthy fish from non-exposed cages at non-infected sites but from a region historically infected (Pop I); (ii) a low prevalence population included apparently healthy fish from exposed cages at declared infected sites (Pop II); (iii) a moderate prevalence population included apparently healthy fish from infected cages after an ISA outbreak (Pop III); and (iv) a high prevalence population included a mixture of apparently healthy, dead and moribund fish from infected cages during an ISA outbreak (Pop IV).

4.2.2 Subject recruitment and sample collection

A total of 400 fish represented multiple origins: 100 from Pop I (50 fish from two different cages, different sites); 130 from Pop II (20, 50 and 60 fish from three different cages, respectively, different sites); 70 from Pop III (20 and 50 fish from two different cages, respectively, different sites); and 100 from Pop IV (10, 14, 31 and 45 fish from four different cages respectively, different sites). Except for Pop IV, salmon were collected using systematic random sampling during the harvest. For Pop IV, 10 fish were slow swimming salmon from a declared infected cage conveniently dipped from the side of the cage; 45 fish (14+ 31) were fresh mortalities and moribund fish collected purposively during an outbreak; and the last 45 fish were sampled randomly during the culling of an outbreak cage.

Following removal from the cage, fish were preserved on ice until kidney samples were taken. For each fish, abdominal organs, including the swim bladder, were removed and the membranous capsule of the kidney was aseptically detached. A series of 1-2 mm cubed sections of the kidney were collected into sterile 2 mL microtubes containing 1 mL of RNA*later* (Ambion Inc., Austin, TX, USA) and the remaining tissue was collected and stored into sterile tissue bags. Microtubes were subsequently stored at -80 °C after a 24 hour period at 4 °C, and tissue bags were directly archived at -80 °C. Multiple impression smears of kidney were made on slides. Slides were heat- and then acetone-fixed for 10 min before being stored at -80 °C. Samples were sent to the testing laboratories on dry ice to ensure that the impact of transport on specimen quality was minimized.

4.2.3 Data collection and management

4.2.3.1 Testing protocols and interpretation

To avoid review bias (Ransohoff & Feinstein, 1978), kidney samples were randomly coded to blind operators. Samples and test allocations for ISAV are illustrated in Fig. 4.1.

4.2.3.1.1 Conventional reverse-transcriptase polymerase chain reaction (RT-PCR)

The assay of interest for this study was a one-step RT-PCR, and the detailed protocol for it was described in Chapter II. From RNA*later*-preserved kidney tissue, the assay detected the presence or absence of a 120 base fragment (analytical target) of the 8th RNA segment of the viral genome. Ultimately, the diagnostic purpose of the assay was to detect infection in the tested fish (presence of active viral particles). Therefore, it was assumed that strong positive correlation exists between the presence of the targeted RNA fragment and the presence of segment 8, the presence of segment 8 and the presence of viral genome, the presence of viral genome and the presence of viral particles (both active or inactive), the presence of viral particles and the presence of active viral particles, and the presence of active viral particle in the kidney sample and the true infection status of the tested salmon (sampling effect as discussed in Thurmond & Johnson, 2004). Evaluation of the operating characteristics of a test includes sampling and transport procedures, the detection process, and interpretation of the test outcomes

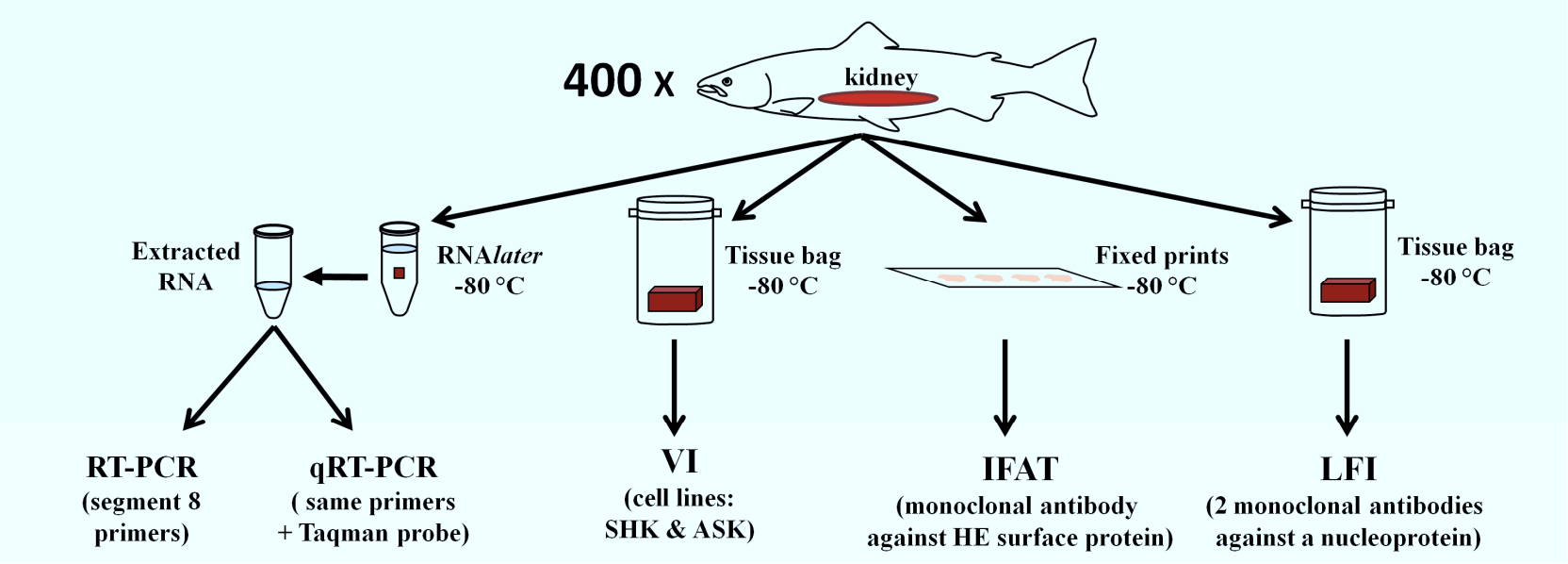


Fig. 4.1. Sample collection & test allocation to evaluate five ISAV detection assays.
Studied assays included: reverse-transcriptase polymerase chain reaction (RT-PCR), real-time RT-PCR (qRT-PCR), virus isolation (VI), indirect fluorescent antibody test (IFAT), lateral flow immunoassay (LFI).

(OIE, 2009b). The test targeted a conserved region of segment 8 and could in theory detect any sub-type of ISAV, regardless of its virulence (i.e. including HPR0).

4.2.3.1.2 Real-time or quantitative Reverse-Transcriptase Polymerase Chain Reaction (qRT-PCR)

The real-time version of the RT-PCR used the same pair of primers in addition to a Taqman probe. The detailed protocol of this assay was described in Chapter III (p.). This assay was performed on the same RNA extracts as conventional RT-PCR (Fig. 4.1), thus the same serial assumptions of positive correlation between the presence of the targeted RNA fragment (analytical target) and infection in the tested fish (diagnostic target) is made. The outcome of the qRT-PCR was given by the amplification cycle at which the fluorescent signal was measured above a specified threshold (cycle threshold, Ct). For any sample that yielded a Ct value before the end of the reaction (45 cycles total), the test result was deemed positive. Using an additional probe, qRT-PCR was believed to be more specific (analytical and diagnostic) than RT-PCR, albeit it also detects any sub-type of ISAV.

4.2.3.1.3 Virus isolation (VI)

The approach to ISAV isolation used in this study was developed by Rolland et al. (2005) and uses two cell lines in parallel. Using frozen kidney samples conditioned in tissue bags, duplicate cell cultures of salmon head kidney 1 (SHK-1) and Atlantic salmon kidney (ASK) were seeded simultaneously. If, within 28 days, at least one of the wells showed cytopathic effects (CPE), the culture was sub-seeded in new duplicate cultures of

both cell lines. If CPE was confirmed in the second passage, the supernatant was tested with an adjunct test (i.e. RT-PCR) for identification. The results of the combined VI were interpreted in parallel (i.e. if at least one cell line tested positive, the sample was deemed positive) (Dohoo et al., 2009). This interpretation in parallel was intended to increase DSe but assumed some degree of conditional dependence between the two VI procedures. This assay targeted the presence or absence of active virulent ISAV particles (analytical target) in the tested sample. To fit the diagnostic target of detecting infection in the tested salmon, it was assumed that a series of strong positive correlations exist between the detection of active viral particles in kidney sample and the true infection status of tested fish. However, VI only detects virulent ISAV genotypes capable of replication in the cell culture (i.e. not HPR0) which may affect the validity of comparing it with RT-PCRs.

4.2.3.1.4 Indirect fluorescent antibody test (IFAT)

The IFAT protocol to detect ISAV from kidney imprints was initially developed by Falk & Dannevig (1995) and was fully described by Nérette et al. (2005). The IFAT uses a monoclonal antibody that binds to the surface glycoprotein hemagglutinin-esterase (HE) of the viral particle (analytical target) (Falk et al., 1998). Similar to other tests, it was necessary to extrapolate the diagnostic target (detection of infection in the tested salmon) from the detection of the HE in the kidney smear by assuming a series of positive correlations. The IFAT results were reported using a grade scoring system according to frequency of observed fluorescent areas (0 to 4+), and any samples that generated at least 1+ score were deemed positive. The monoclonal antibody used has

been shown to bind to all of the North American genotypes of ISAV known, and there is no evidence suggesting that the antibody does not bind to the avirulent genotype HPR0 (Dr. Knut Falk, pers. com.).

4.2.3.1.5 Lateral flow immunoassay (LFI)

The LFI is a commercially available assay (Aquatic Diagnostics Ltd., Stirling, Scotland) that uses a mix of two monoclonal antibodies targeting the nucleoprotein (analytical target) coded by the 3rd RNA segment of the viral genome. Although the test was developed for fresh kidney samples, frozen tissues conditioned in tissue bags were used for this evaluation. A series of positive correlations were assumed between the presence of the nucleoprotein and the ultimate diagnostic target. Both monoclonal antibodies are believed to bind to all North American genotypes of ISAV and, no evidence suggests that they do not bind to the avirulent genotype HPR0 (Dr. Kim Thompson, pers. com.).

4.2.3.2 Data alignment and test agreement

Visual screening and agreement of the 5 assay results was conducted using the descriptive approach developed in Chapter II. Separate alignment of the test results was generated (fish in column, test in row) for each population using the DNA sequence alignment editor software BioEdit version 7.07 (Hall, 1999). Agreement among the 5 tests was represented with an agreement tree reconstructed using the phylogenetic

analysis package MEGA version 4 (Tamura et al., 2007). Analysis parameters for the tree reconstruction are detailed in Chapter II.

4.2.3.3 Suspected conditional dependence

Diagnostic tests that target a similar biologic trait (i.e. same analyte and/or stage of the infection) are expected to be dependent conditional on the infection status (Gardner et al., 2000). Therefore, the two nucleic-acid amplification tests (NAAT) (i.e. RT-PCR & qRT-PCR) and the two antibody based assays (ABA) (i.e. IFAT & LFI) were expected to be conditionally dependent. Although these two pairs of tests were particularly targeted, statistical procedures were used to investigate test dependence for any test pairs.

4.2.4 Latent Class Modelling

Allowing for more flexibility, LCMs were coded and run in a Bayesian framework using WinBugs software (Spiegelhalter et al., 2003). The code for the final model is available in Appendix 17.

4.2.4.1 Parameters and identifiability

Test results were collected from $P = 5$ assays for each of the $N = 400$ sampled salmon clustered in $K = 4$ different prevalence populations. To investigate the assumption of constant DSe across populations and test conditional independence, we considered $C =$

3 categories (or latent classes) of infection: the class of non-infected (L_A), and two sub-classes of infected fish (L_B and L_C). L_B is assumed to be an intermediate stage between L_A and L_C. Similar to Dendukuri et al. (2009), we set a model where Y_{ip} denotes the result of the i^{th} fish on the p^{th} test, and $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5})$ and denotes the result vector for the i^{th} salmon. Y_{ip} is dichotomous (negative or positive; 0 or 1, respectively). Assuming that the 5 tests are conditionally independent within each latent class, the probability to observe Y_i in population k was expressed as:

$$P_k(Y_i) = \sum_{c=1}^{3} P_k(L_c)\, P_k(Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5}|L_c)$$

$$= \sum_{c=1}^{3} P_k(L_c) \prod_{p=1}^{5} P_k(Y_{ip}|L_c) \qquad (1)$$

where $P_k(L)$ denotes the prevalence of class L in the k^{th} population; and $P_k(Y_{ip}|L_c)$ denotes the probability that the i^{th} fish yields a result Y_{ip} (0 or 1) on the p^{th} test in class c of the k^{th} population. In the two classes of infected fish (L_B, L_C), the probability to test positive correspond to the DSe, while in the class of non-infected (L_A), it corresponds to the complement of DSp (1-DSp). These parameters were assumed to be population independent and therefore constant across the 4 populations. In each population, the probability of non-infected fish ($P(L_A)$) can be expressed as the complement of the prevalence of infection (i.e. $1- (P(L_B)+P(L_C))$). Within each population k, the result vector Y_i is a count that follows an independent multinomial sampling distribution.

Under the assumption of independence of test results within each of the 3 classes, 23 unknown parameters were included in the model ($CP + K(C-1)$, modified from Dendukuri et al., 2009); in detail, one DSp and two DSe (DSe_{LB} and DSe_{LC}) for each of

the 5 tests, and two infection prevalences (L_B and L_C) in each of the 4 populations. The dataset degrees of freedom were 124 (K (2^P-1); Hui & Walter, 1980).

4.2.4.2 Prior information

Non-informative beta priors (1, 1) were set for DSp and DSe. Prevalence of the 3 classes in each population were set to follow a Dirichlet distribution (whereby the 3 proportions add to 1) and were parametrized using 3 gamma distributions, as described in the WinBugs manual (Spiegelhalter et al., 2003). Initially, the gamma distributions were given the same shape parameters (1, 1, 1), leading to a non-informative Dirichlet prior, in all 4 populations. However, to avoid identity switching between the two sub-classes of infected, a weak prior distribution was set for the prevalence of L_A fish in Pop I. In a previous ISAV prevalence study, the proportion of non-infected fish (class A) in non-exposed cages in an infected area was estimated at 95% (McClure et al., 2004). Using the BetaBuster freeware (http://www.epi.ucdavis.edu/diagnostictests), distribution parameters were calculated for a prevalence distribution corresponding to a mode at 95% and a 5% percentile at 50% (i.e. beta(4.8, 1.2)). The beta prior was converted into a Dirichlet prior by distributing the non-class A proportions equally on classes B and C (i.e. gamma(4.8, 0.6, 0.6)).

4.2.4.3 Model refinement for conditional dependence among tests

Although conditional dependence was assumed to be addressed by adding an extra class, test dependence within each class was further explored, one test pair at a time, using the methods described by Nérette et al. (2008a). Models corresponding to the 10 possible combinations of test pairs were compared using the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and Bayesian *P*-values (Nérette et al., 2008a). DIC reflects the combination of 2 parameters: *pD*+ describing the model complexity (i.e. number of effective parameters), and *D* describing the goodness-of-fit. Models were considered significantly different when DIC differed by more than 3 units. In addition, Bayes *P*-values were estimated to select models based on their goodness-of-fit using a code adapted from Nérette et al. (2008a). Only covariance terms between the two NAATs and the two ABAs were significant and included in the final model. Further model refinements were pursued and only significant covariance terms conditional on the infection class C were conserved in the final code. From Eq. (1), the probability that the i^{th} fish yields a result Y_{ip} (0 or 1) on the p^{th} test in the class C included the covariance terms and was expressed as follow:

$$P(Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5}|L_c) = P(Y_{i1}, Y_{i2}|L_c) * P(Y_{i3}|L_c) * P(Y_{i4}, Y_{i5}|L_c)$$

When $Y_{i1}=Y_{i2}$,

$$P(Y_{i1}, Y_{i2}|L_c) = P(Y_{i1}|L_c) * P(Y_{i2}|L_c) + \gamma_{12}$$

and when $Y_{i1} \neq Y_{i2}$,

$$P(Y_{i1}, Y_{i2}|L_c) = P(Y_{i1}|L_c) * P(Y_{i2}|L_c) - \gamma_{12}$$

where γ_{12} is the covariance term between test 1 and 2 (NAATs) in class C. Similar formulae apply to the conditional probabilities involving tests 4 and 5 (ABAs) ($P(Y_{i4}, Y_{i5}|L_c)$), with a covariance term (γ_{45}) between tests 4 and 5 in class C.

4.2.4.4 Assessment of MCMC chains convergence

For all models, a burn-in period of 10,000 iterations and 40,000 additional iterations were run for posterior sampling estimation. Proper convergence and mixing of Markov Chain Monte Carlo were assessed by following Toft et al. (2007) guidelines. A brief overview of the steps in this process includes the following: (i) visual assessment of time series trace plots using Gelmann (1996) empiric criteria; (ii) Gelman-Rubin convergence diagnostics (Brooks & Gelman, 1998) to assess the influence of the starting values on the chains convergence; and (iii) autocorrelation plots to assess the presence of correlation among posterior samples. To reduce some observed autocorrelation along the Gibbs sampler chains, a total of 400,000 iterations were run with thinning sampling every 10 iterations to ultimately obtain 40,000 posterior samples. In addition, chain convergence assessment was complemented with quantile stability diagnostics (Raftery-Lewis, 2.5% and 97.5%; Brooks-Draper, mean) and the Effective Sample Size (i.e. sample size for a completely uncorrelated sequence that would yield the estimated MC standard error for the mean) using the "column diagnostic" function in MLwiN software v.2.11 (Rasbash et al., 2009). Markov chains were run in parallel using three random sets

of initial values for test parameters (i.e. DSe_A, DSe_B and DSp). Some convergences failed due to a shift between the two sub-classes of infected fish relative to the initial values. New sets of initial values were arbitrarily set based on the overall interpretation of the two sub-classes of infected fish (i.e. DSe_B lower than DSe_A).

4.2.4.5 Model validity

The goodness-of-fit of the model was assessed using the Bayesian *P*-values as described in Nérette et al. (2008a). In addition, the robustness of the model estimation was investigated with respect to the influence of various parameters on the analysis. First, the influence of the third class on the LCM procedure was assessed by running a conventional two-class LCM (2LCM) for comparison. Next, the influence of conditional dependence between NAATs and ABAs was studied by breaking the dependent pair of tests and running 2LCM and 3LCM models with 3 tests only, including combinations using one NAAT and one ABA simultaneously. Covariance factors were conserved when the selected pair of tests was present in the model.

To investigate the potential impact of a particular test, the model was run with four tests at the same time. Robustness of the model was evaluated by monitoring the variation of estimated parameters across different cut-point values for ordinal or continuous outcome tests. The analysis was repeated for different qRT-PCR Ct cut-off value and IFAT score cut-off values (1, 2, 3, 4). Finally, the influence of the informative prior (prevalence of class A in Pop I) was explored by running the model with first a non-informative prior (gamma(1, 1, 1)), or with a weaker prior distribution (gamma(1.3, 0.5,

0.5)) with the mode at 95% and the 5% percentile at 10%), or finally with a stronger prior distribution (gamma(10.8, 0.75, 0.75)) with the mode at 95% and the 5% percentile at 70%) (Fig. 4.2).

4.3 Results

4.3.1 Test result alignment and agreement tree

For each population, the test results were aligned and were presented in Fig. 4.3. Three different testing patterns were observed. The first pattern was fish yielding negative results to all 5 tests and was found in all populations, but largely dominated in Pop I and II. The second testing pattern was salmon with at least 3 positive tests and was exclusively found in Pop IV (infected cages). The third pattern was observed mainly in Pop III, and included fish positive to either both NAATs and negative in the 3 other methods. The observation of 3 distinct testing patterns supported the addition of a third class in the analysis. A table of test result frequencies is presented in Appendix 15.

The distance-based tree reconstruction generated a visual representation of the agreement among the 5 assays (Fig. 4.4). The tree is scaled with the number of test results that differ between assays. As the branch distance between two tests becomes greater, the number of disagreeing results becomes more important. The two antibody-based assays (ABAs) clustered together and revealed the best agreement with only 8 discrepancies out of the 400 fish (2%). The two nucleic-acid amplification tests (NAATs) also clustered together. However, they revealed much stronger disagreement (9.25%),

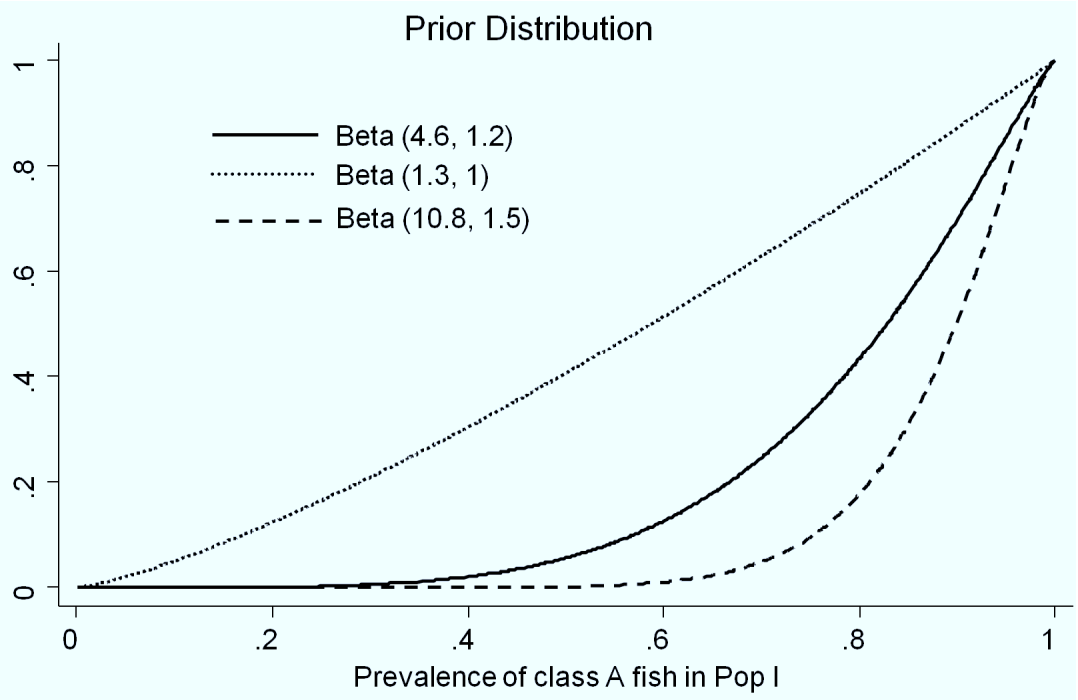


Fig. 4.2. Prior distributions for proportion of class A salmon (non-infected) in the low prevalence population (Pop I).
The distribution used was beta(4.6, 1.2) (mode at 0.95 and the 5% percentile at 0.50). Alternative priors were tested: a beta(1.3,1) (mode at 0.95 and the 5% percentile at 0.10); and a beta(10.8, 1.5) (mode at 0.95 and the 5% percentile at 0.70).

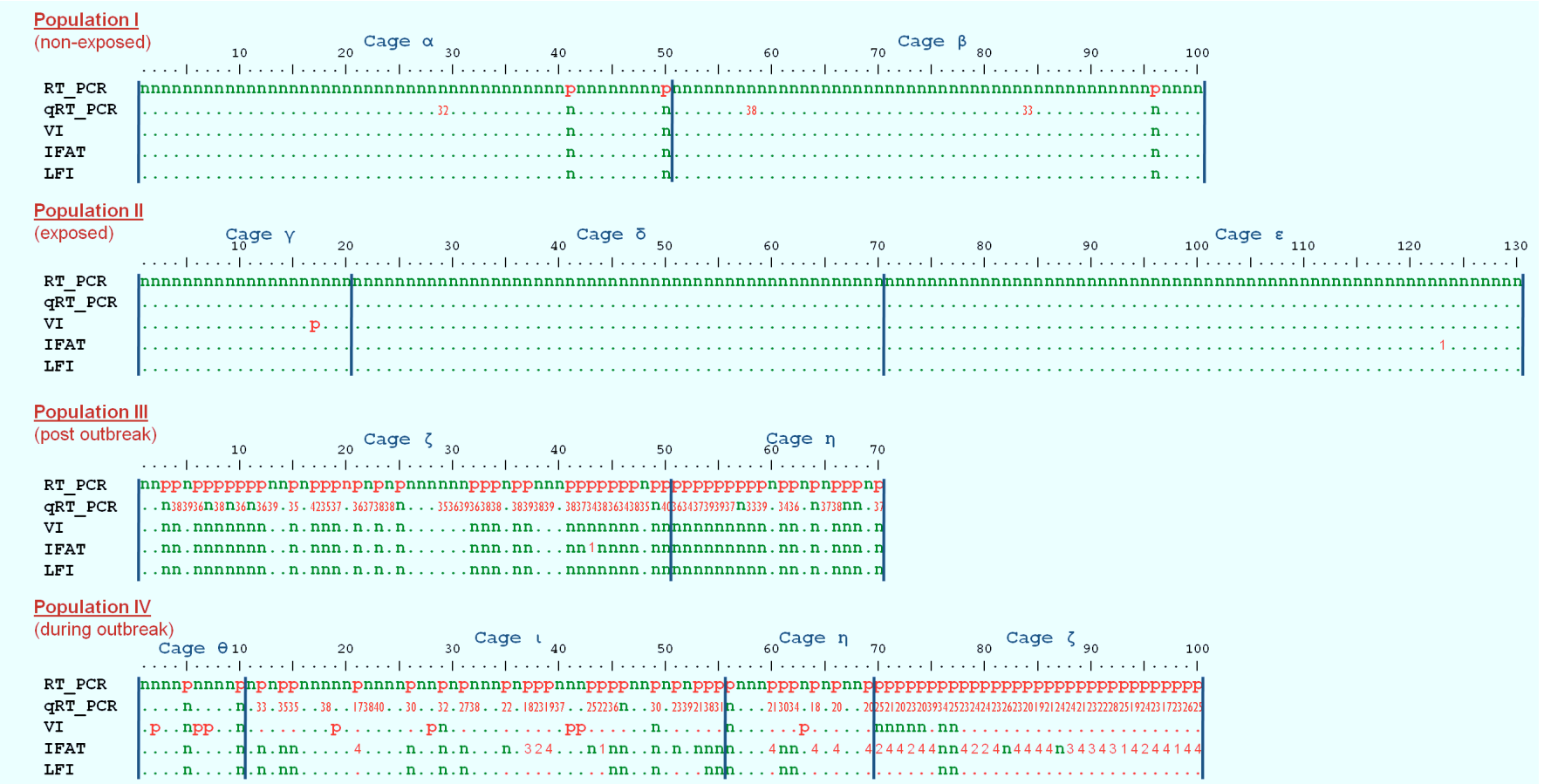


Fig. 4.3. Test result alignment: sampled fish (in columns) were clustered by cage origin and prevalence level populations (Population I: 100 apparently healthy fish from non exposed cages; Population II: 130 apparently healthy fish from exposed cages; Population III: 70 apparently healthy fish from post outbreak cages; Population IV: mixture of apparently healthy, mortality and moribund fish from outbreak cages). RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay; "n" codes for negative; "p" codes for positive; a dot indicates the same result as the first row (RT-PCR). When positive, cycle threshold values were reported for qRT-PCR, and intensity score (0 to 4) for IFAT. Greek letters are arbitrary cage numbers. Note that the same cages were sampled for moderate and high prevalence populations.

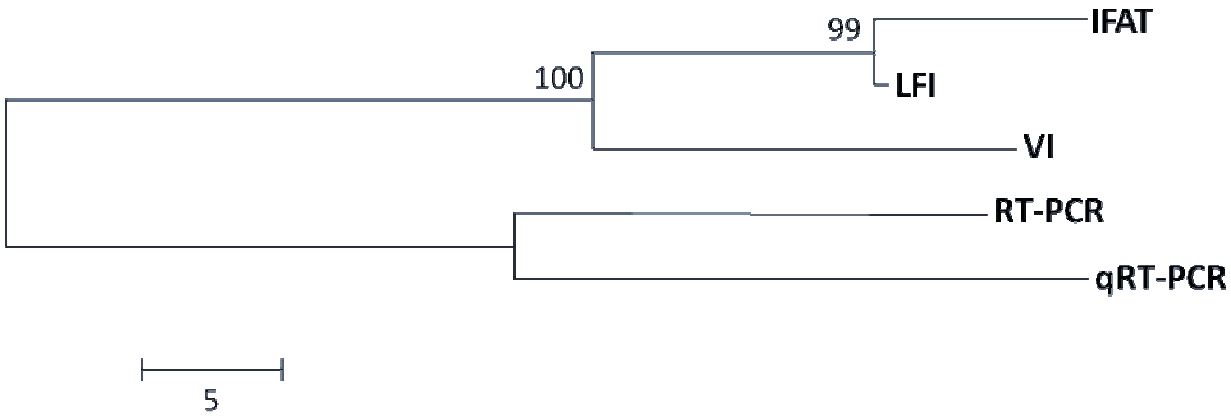


Fig. 4.4. Unrooted phylogram representing agreement among test runs.

The percentages reported at the branch node are the bootstrap support values (proportion of resampled trees that include the node of interest) and the distance between two tests is visually assessed by the relative length of branches that connect them and are scaled based on the number of differing results out of the 400 samples tested. RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFI: lateral flow immunoassay.

particularly in Pop III. The VI branched with the ABAs cluster, suggesting a similar testing pattern but sufficiently different to be separated. The respective cluster of the 2 ABAs and the 2 NAATs suggested that some degree of dependency was to be expected within each these pairs of tests.

4.3.2 Model building and refinement

Compared to the model with conditional independence, significant improvement in the model DIC and goodness-of-fit were observed when dependence between the 2 NAATs and the 2 ABAs were accounted for (Table 4.1). Further refinement of the model revealed that dependence within the two pairs of tests was only conditional on class C and covariance terms in the two other classes could be excluded without changing substantively the model's DIC (Table 4.2).

4.3.3 Final three-class LCM (3LCM)

The final model included 3 classes with 400 salmon from 4 different populations tested by 5 tests and accounting for test dependence between the NAATs and the ABAs, and required thinning. The model showed satisfactory goodness-of-fit (Bayesian *P*-value = 0.56) and a DIC of 145.95. Posterior distributions of the test result probabilities were presented in separate graphs for each class (Fig. 4.5A, B & C), and detailed estimates from the distributions (mean, median and mode) and their corresponding 95% credibility intervals were reported in Table 4.3. For all tests, the probability to test negative in class A (DSp) was substantial, exceeding 98%. LFI had the highest DSp and small variation of

Table 4.1
Model selection directed by Bayesian *P*-value and deviance information criterion (DIC) to identify the conditional dependence among pairs of tests.
Best model in bold. RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay.

Model	RT-PCR	qRT-PCR	VI	IFAT	LFI	Bayes-*P*	DIC
I						0.23	204.16
II	•	•				0.26	179.66
II	•		•			0.23	181.62
IV	•			•		0.23	183.17
V	•				•	0.24	181.38
VI		•	•			0.24	182.70
VII		•		•		0.23	182.86
VIII		•			•	0.24	180.47
IX			•	•		0.24	181.83
X			•		•	0.25	181.21
XI				•	•	0.42	152.01
XII	•	•		•	•	**0.55**	**148.47**

Table 4.2
Model selection guided by Bayesian *P*-values and deviance information criterion values (DIC) **to identify the best combination of covariance factors between the 2 nucleic acid amplification tests** (NAAT: RT-PCR & qRT-PCR) **and between the two antibody-based assays** (ABA: IFAT & LFI), **conditional on the infection status.** Best model (most parsimonious) in bold. RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay.

Model	covA-NAAT	covA-ABA	covB-NAAT	covB-ABA	covC-NAAT	covC-ABA	Bayes-*P*	DIC
I	•	•	•	•	•	•	0.55	148.47
II		•	•	•	•	•	0.58	146.92
II	•		•	•	•	•	0.55	148.07
IV	•	•		•	•	•	0.55	147.96
V	•	•	•		•	•	0.55	148.16
VI	•	•	•	•		•	0.40	154.07
VII	•	•	•	•	•		0.28	180.42
VIII					•	•	**0.54**	**146.17**
IX					•		0.25	181.37
X						•	0.40	151.44
XI							0.23	204.16

covA: covariances for class A (non-infected) (specificity)
covB: covariances for class B (infected) (sensitivity)
covC: covariances for class C (infected) (sensitivity)

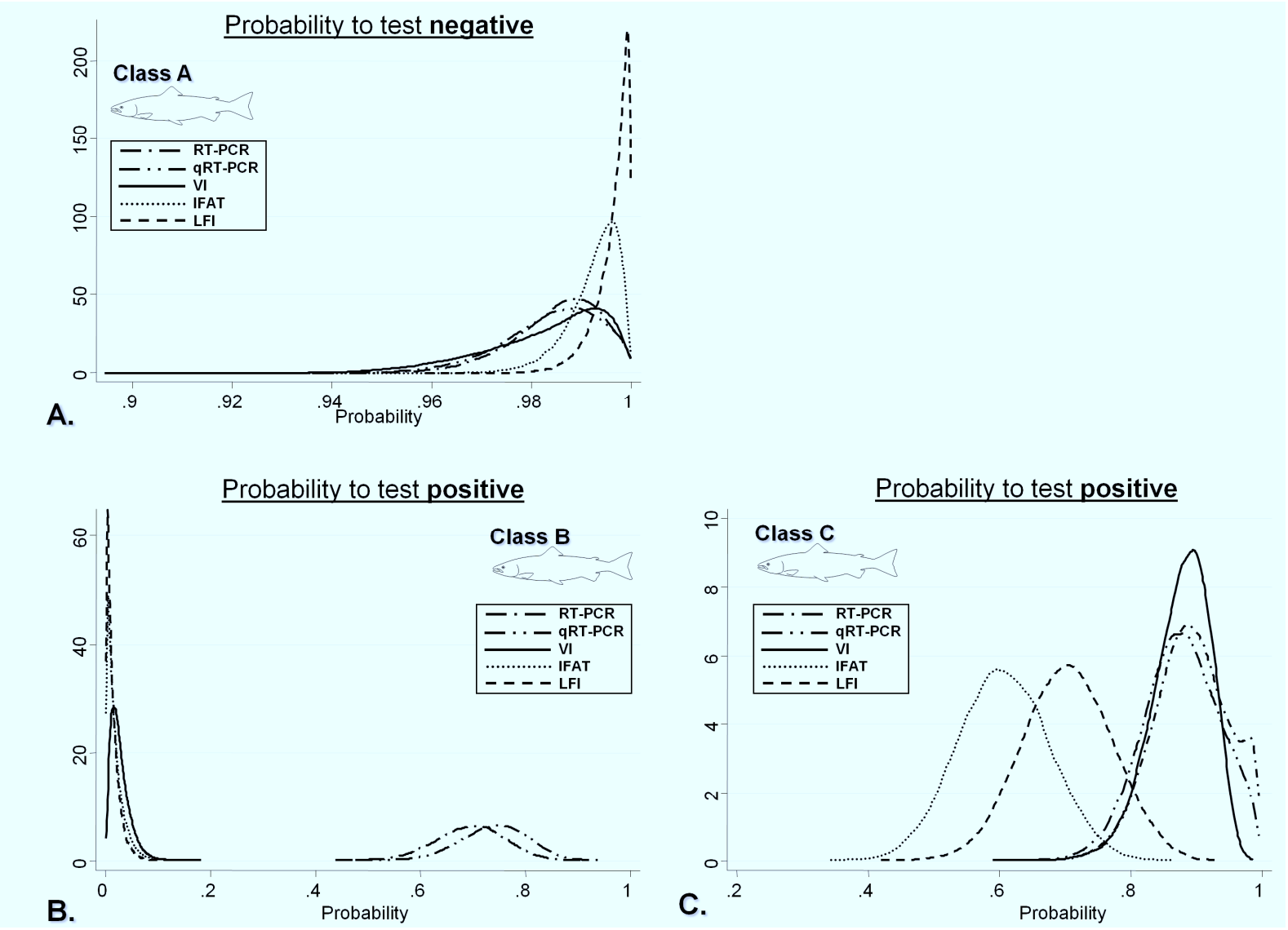


Fig. 4.5. Posterior distributions of the probabilities of each ISAV assay to test positive if the fish is infected in class A: DSe_A (A.); infected B: DSe_B (B.); and non-infected: 1-DSp (C.).

RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay; DSe: diagnostic sensitivity; DSp: diagnostic specificity.

Table 4.3

Posterior estimates (mean, median and mode) **and corresponding 95% credibility posterior intervals** (CPI) **of probabilities of testing positive and negative in the three class of fish for each of the five ISAV diagnostic assays and class prevalences for each of the 4 populations from the final 3LCM, including conditional dependence.** RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay; DSe: diagnostic sensitivity; DSp: diagnostic specificity; 3 classes of fish (A, B & C) where B & C are assumed infected.

Estimates (%)	**RT-PCR**				**qRT-PCR**				**VI**				**IFAT**				**LFI**			
	Mean	*Median*	*Mode*	*95% CPI*	*Mean*	*Median*	*Mode*	*95% CPI*	*Mean*	*Median*	*Mode*	*95% CPI*	*Mean*	*Median*	*Mode*	*95% CPI*	*Mean*	*Median*	*Mode*	*95%CPI*
Class A																				
Prob (negative) DSp	98.51	98.65	98.97	96.4-99.9	98.39	98.53	98.81	96.0-99.9	98.27	98.58	99.29	95.2-99.8	99.25	99.37	99.66	97.9-99.9	99.63	99.74	99.97	98.6-99.9
Class B																				
Prob (positive) DSe	69.32	69.52	70.66	56.4-81.3	73.85	74.12	71.26	61.0-85.2	1.70	1.19	1.01	0.04-6.16	2.60	2.19	1.37	0.31-7.21	1.31	0.91	1.13	0.03-4.76
Prob (negative) DSp	30.68	30.48	29.34	18.7-43.6	26.15	25.88	28.74	14.8-39.0	98.30	98.81	98.99	93.8-99.9	97.40	97.81	98.63	92.8-99.7	98.69	99.09	98.87	95.2-99.9
Class C																				
Prob (positive) DSe	89.56	89.60	89.78	78.2-99.4	88.10	88.18	88.37	76.4-98.5	88.13	88.60	90.20	78.3-95.3	61.07	61.00	62.20	47.6-74.7	70.81	70.83	71.67	57.4-84.0
Covariance*				7.37 7.58 10.85	0.14-15.1											15.91 16.16	16.81	9.71-20.7		

Prevalence (%)	**Population I**				**Population II**				**Population III**				**Population IV**			
	Mean	*Median*	*Mode*	*95% CPI*	*Mean*	*Median*	*Mode*	*95% CPI*	*Mean*	*Median*	*Mode*	*95% CPI*	*Mean*	*Median*	*Mode*	*95% CPI*
Class A	97.3	97.9	96.8	91.2-99.9	98.4	98.6	97.2	95.6-99.8	12.8	12.6	9.2	2.36-24.5	29.2	28.9	27.4	19.0-40.9
Class B	2.16	1.41	1.13	0.01-8.05	0.82	0.57	1.01	0.02-3.00	85.8	85.9	84.6	73.7-96.5	12.3	11.9	10.9	5.14-20.9
Class C	0.57	0.30	1.13	0.00-2.64	0.78	0.55	1.03	0.02-2.87	1.44	1.00	1.14	0.03-5.29	58.5	58.7	59.6	46.9-69.2

*Conditional dependence between the 2 RT-PCRs and between the 2 antibody-based tests

estimation (narrow distribution), while the DSp of VI was comparable to the NAATs but showed the largest uncertainty in the estimation (i.e. lack of evidence). In class B, only NAATs were likely to test positive while the other assays had almost zero probability. In class C, the probability for the ABAs to test positive was substantially lower compared to the three other tests. The LFI seemed to perform better than the IFAT that had the lowest DSe. VI has the highest DSe mode, albeit its narrow distribution suggested limited chance to reach higher values than 95%. Although the distribution of DSe for RT-PCR was wider, its average DSe estimate was the highest, as explained by a secondary peak (Fig. 4.5C). The DSe of qRT-PCR distribution also showed a tendency to bimodality, despite its profile appearing symmetrical. Significant dependencies between the NAATs and the ABAs were confirmed conditional on class C (Table 4.3).

Posterior distributions of the prevalence of class B and C (infected fish) were presented in separate graphs for each population (Fig. 4.6A, B, C & D); and the detailed estimates were reported in Table 4.3. The prevalence of infected fish were almost nil in Pop I and Pop II. A Class B of infected fish was largely present in Pop III (post-outbreak cages). Class B infected fish were also found in a minority of cages encountering an outbreak (Pop IV), dominated by the presence of Class C fish.

4.3.4 Sensitivity analyses

Conventional 2LCM with 5 tests showed strong evidence of lack of fit (Bayesian *P*-value = 0.01) (Table 4.4). Accounting for conditional dependence between the NAATs and ABAs did not particularly improve the fitness (Bayesian *P*-value = 0.03) (Table 4.4).

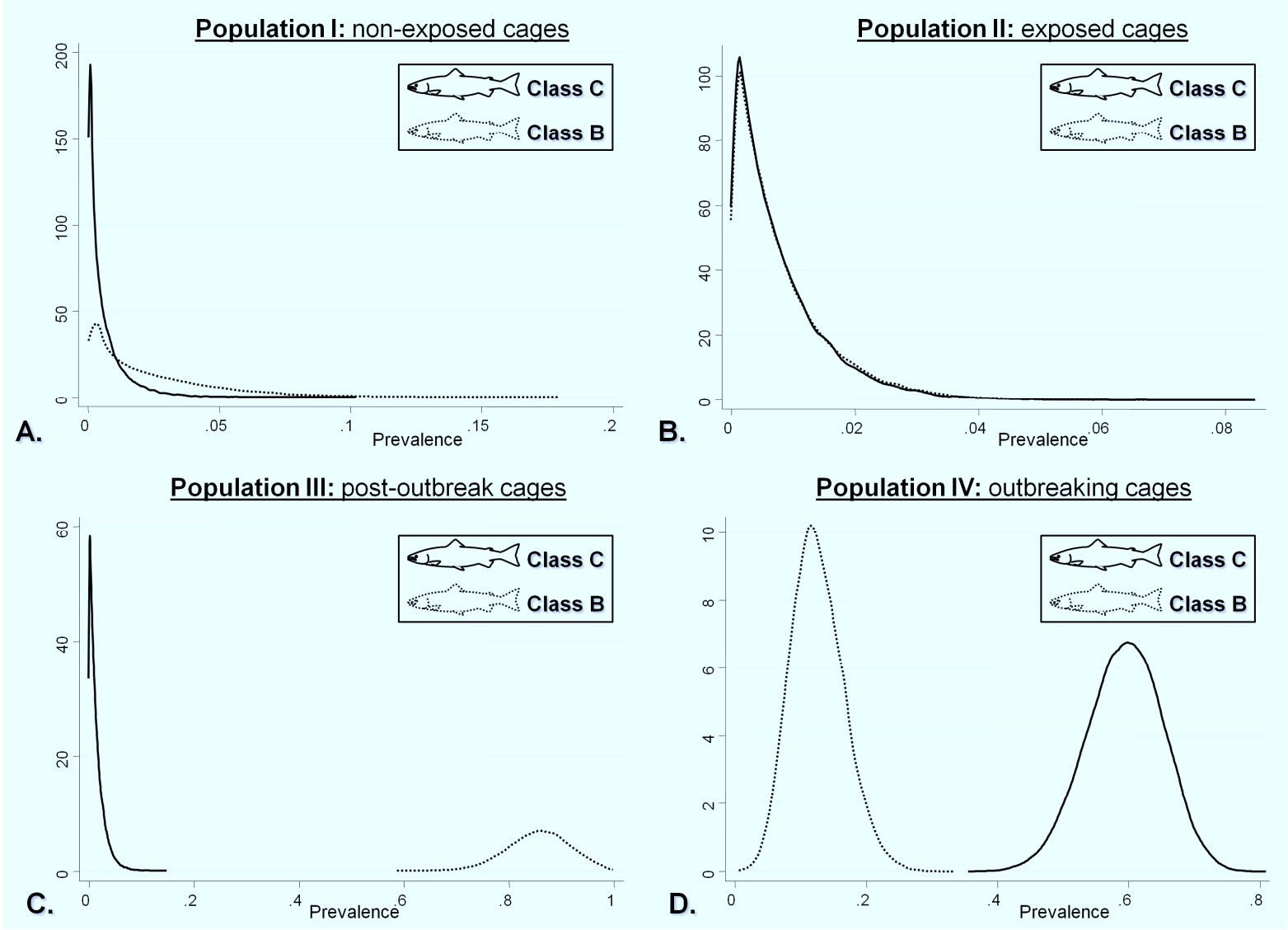


Fig. 4.6. Posterior distributions of the prevalences of class B and C of each of the 4 sampled populations: Population I (fish from non-exposed cages, low prevalence expected), Population II (fish from exposed cages, mild prevalence expected), Population III (fish from post-outbreak cages, moderate prevalence expected), Population IV (fish from cages encountering an outbreak, high prevalence expected).

Table 4.4

Posterior means of 2- and 3-class models with and without test conditional dependence (CD).

RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay; DSe: diagnostic sensitivity; DSp: diagnostic specificity; 3 classes of fish (A, B & C) where B & C are assumed infected.

Model	*covA-NAAT*	*covA-ABA*	*covB-NAAT*	*covB-ABA*	*covC-NAAT*	*covC-ABA*	RT-PCR			qRT-PCR			VI			IFAT			LFI			Pop I		Pop II		Pop III		Pop IV		*Bayes-P*	*DIC*
							DSp	*DSe B*	*DSe C*	*DSp*	*DSe B*	*DSe C*	*DSp*	*DSe B*	*DSe C*	*DSp*	*DSe B*	*DSe C*	*DSp*	*DSe B*	*DSe C*	*Prev. B*	*Prev. C*	*Prev. B*	*Prev. C*	*Prev. B*	*Prev. C*	*Prev. B*	*Prev. C*		
2LCM	-	-	-	-	-	-	95.0	-	96.7	93.6	-	95.5	96.6	-	49.5	99.3	-	38.4	99.6	-	43.5	-	1.00	-	0.75	-	55.2	-	58.1	0.01	433.43
2LCM-CD	0.28	0.18	-	-	4.47	17.9	97.6	-	80.4	97.4	-	82.3	98.6	-	42.5	99.1	-	29.1	99.5	-	32.9	-	1.27	-	0.81	-	78.1		72.4	0.03	339.61
3LCM	-	-	-	-	-	-	98.5	69.5	98.0	98.4	73.2	97.4	96.4	6.89	87.8	99.1	3.13	72.4	99.6	1.22	84.0	2.10	0.57	0.82	0.75	84.1	1.38	17.6	48.9	0.23	204.16
3LCM-CD	0.25	0.18	0.30	0.64	7.13	1.58	98.5	68.4	89.8	98.4	73.4	88.3	98.2	1.70	88.2	99.1	3.09	61.3	99.5	1.49	71.0	2.89	0.56	0.81	0.77	86.0	1.42	12.8	58.2	0.54	148.47

The addition of a third class without covariance terms substantially increased the Bayesian *P*-value (0.23), while the best goodness-of-fit (0.54) and lowest DIC were obtained by including covariance factors in each of the 3 classes (Table 4.4).

Alternative ways to avoid conditional dependence using 2LCM were investigated using chosen combinations of 3 tests to break conditional dependence (Table 4.5). Although DSp estimates were similar to the final 5-test 3LCM accounting for conditional dependence, DSe estimates for non-NAAT tests were approximately average for the class B and class C estimates (Table 4.5). Addition of the third class in 3-test models resulted in some unstable estimates for NAATs and the class B prevalence (models VI & VII, Table 4.5). Removing one test at a time (i.e. 4-test 3LCM) did not reveal any substantial changes in the estimation (Table 4.5). None of the tests had a major influence on the model. However, the absence of VI seemed to increase the DSe estimates in class C of other tests (model XI, Table 4.5).

Using different cutpoints for IFAT did not influence the estimation of the parameters (results not shown) (Fig. 4.7), while cutpoints for Ct values between 25 and 30 cycles revealed instabilities mainly related to VI classification (Fig. 4.8). In Pop IV, VI generated positive results that were negative for all other tests, which explained its relatively low DSp. Conversely, some samples tested positive with any other tests were negative on VI, which explained its relatively low DSe in class C fish (Fig. 4.3). Between 25 and 30 Ct cutoff, the data were weaker and the model seemed to give more weight to VI classification. Some of these samples (initially A and C) were reclassified as class B fish (increased proportion of class B and decreased proportions of class A and C in Pop

Table 4.5

Posterior means of models using different combinations of 3- and 4-test models for comparison with the final 5-test model.

RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay; DSe: diagnostic sensitivity; DSp: diagnostic specificity; 3 classes of fish (A, B & C) where B & C are assumed infected.

Model	RT-PCR	qT-PCR	VI	IFAT	LFI	RT-PCR			qRT-PCR			VI			IFAT			LFI			Pop I		Pop II		Pop III		Pop IV		*Bayes-P*	*DIC*
						DSp	*DSe B*	*DSe C*	*DSp*	*DSe B*	*DSe C*	*DSp*	*DSe B*	*DSe C*	*DSp*	*DSe B*	*DSe C*	*DSp*	*DSe B*	*DSe C*	*Prev. B*	*Prev. C*	*Prev. B*	*Prev. C*	*Prev. B*	*Prev. C*	*Prev. B*	*Prev. C*		
I[a]	•		•	•		98.5	-	89.6	-	-	-	98.9	-	46.8	99.3	-	31.9	-	-	-	-	2.11	-	0.85	-	66.3	-	68.9	0.00	213.44
II[a]	•		•		•	98.4	-	90.4	-	-	-	98.7	-	47.1	-	-	-	99.6	-	36.5	-	1.97	-	0.81	-	65.5	-	68.2	0.00	235.05
II[a]		•	•	•		-	-	-	98.6	-	88.1	98.9	-	44.8	99.3	-	30.4	-	-	-	-	2.18	-	0.87	-	68.5	-	72.7	0.00	212.38
IV[a]		•	•		•	-	-	-	98.4	-	88.9	98.8	-	45.2	-	-	-	99.6	-	34.8	-	2.04	-	0.82	-	67.7	-	72.0	0.00	235.69
V	•		•	•		99.2	90.7	87.1	-	-	-	98.9	0.18	84.4	99.3	0.37	55.7	-	-	-	3.51	1.03	0.83	0.82	65.9	0.91	3.82	60.9	0.29	82.19
VI	•		•		•	34.7	0.66	88.1	-	-	-	98.0	1.28	88.4	-	-	-	98.5	0.38	68.9	93.9	1.01	98.1	0.77	5.49	0.84	26.2	57.3	0.16	84.22
VII		•	•	•		-	-	-	32.3	0.69	84.9	98.2	0.99	87.6	97.0	0.77	57.2	-	-	-	94.2	1.01	98.0	0.83	6.03	0.90	21.8	58.7	0.22	87.06
VIII		•	•		•	-	-	-	99.2	91.2	86.5	98.7	1.83	86.0	-	-	-	99.6	1.86	66.8	3.49	1.02	0.82	0.78	67.2	0.81	8.01	59.1	0.14	86.61
IX		•	•	•	•	-	-	-	99.1	80.6	86.3	98.7	1.77	86.4	99.1	3.45	58.3	99.6	1.68	67.6	3.07	0.60	0.95	0.78	80.2	1.43	10.4	61.1	0.44	104.11
X	•		•	•	•	99.1	80.0	88.4	-	-	-	98.6	1.77	86.5	99.1	3.65	58.9	99.6	1.73	68.3	3.14	0.59	0.98	0.78	79.3	1.42	6.87	60.7	0.49	100.26
XI	•	•		•	•	98.5	71.0	97.2	98.5	74.7	97.2	-	-	-	99.1	2.67	80.2	99.6	1.49	93.4	2.80	0.56	0.82	0.75	85.7	1.54	23.5	43.6	0.75	96.26
XII	•	•	•		•	98.5	68.1	89.3	98.5	73.1	87.8	98.3	1.65	88.3	-	-	-	99.6	1.43	69.7	3.06	0.57	0.84	0.77	86.4	1.43	12.9	58.7	0.25	128.97
XIII	•	•	•	•		98.5	68.2	86.9	98.4	73.3	85.4	98.8	1.57	87.8	99.2	2.77	58.0	-	-	-	2.93	0.57	0.84	0.81	85.9	1.64	12.7	60.4	0.36	129.71
XIV*	•	•	•	•	•	98.5	69.3	90.3	98.4	73.8	88.8	98.2	1.72	88.1	99.2	2.61	61.5	99.6	1.31	71.3	2.12	0.57	0.82	0.76	85.7	1.44	12.2	58.2	0.54	146.17

[a] 3-test model run with only 2 classes (i.e. no class B value)

DIC: deviance information criterion

* final model run without thinning (i.e. different DIC and Bayes-*P*)

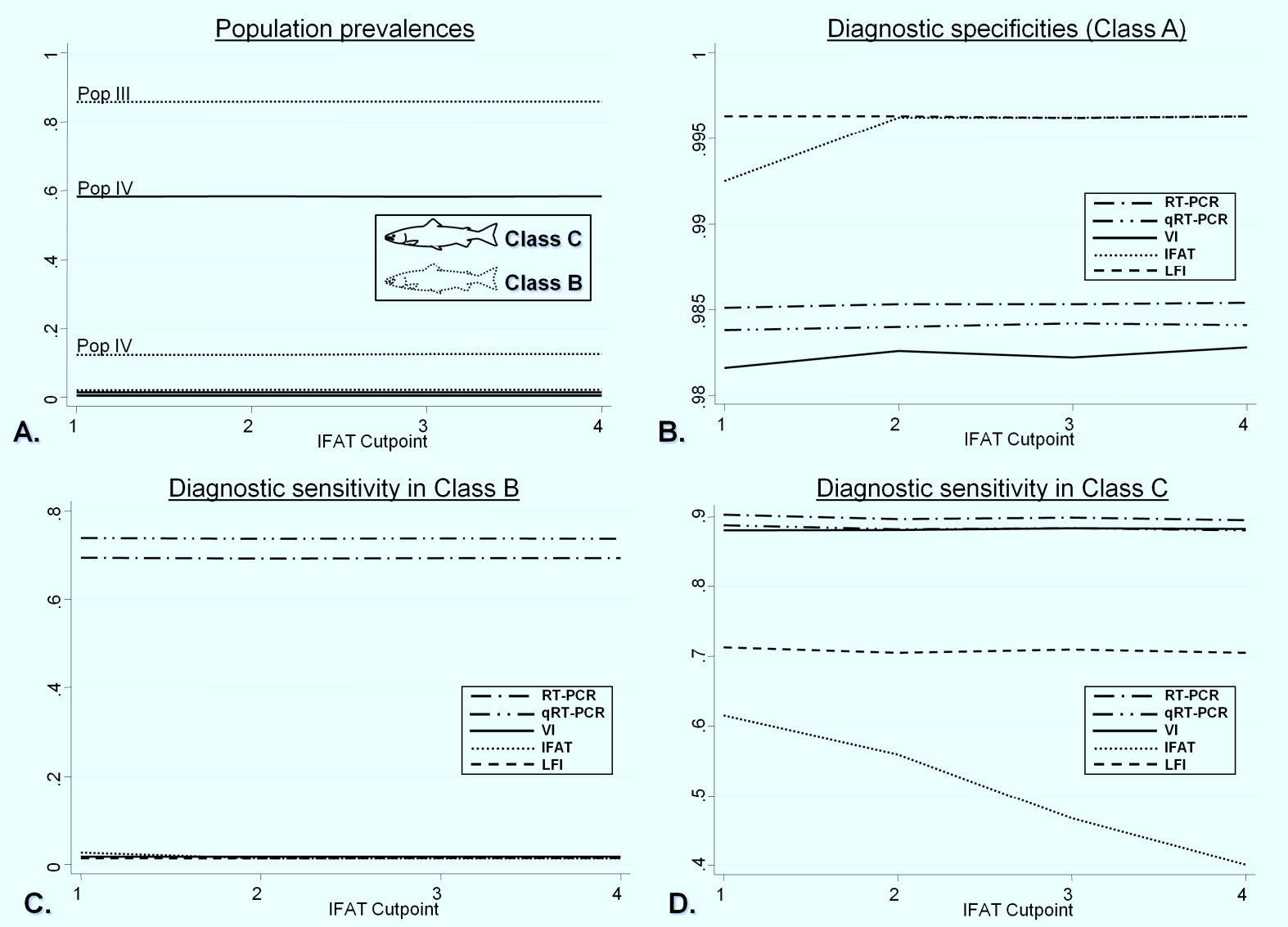


Fig. 4.7. 3-class LCM posterior means of prevalences and test operating characteristics across IFAT cutpoints.
IFAT scoring ranged from 0 to 4. IFAT cutpoint was defined as any samples that yielded a score below the specified value were deemed negative. RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay; DSe: diagnostic sensitivity; DSp: diagnostic specificity; 3 classes of fish (A, B & C) where B & C are assumed infected.

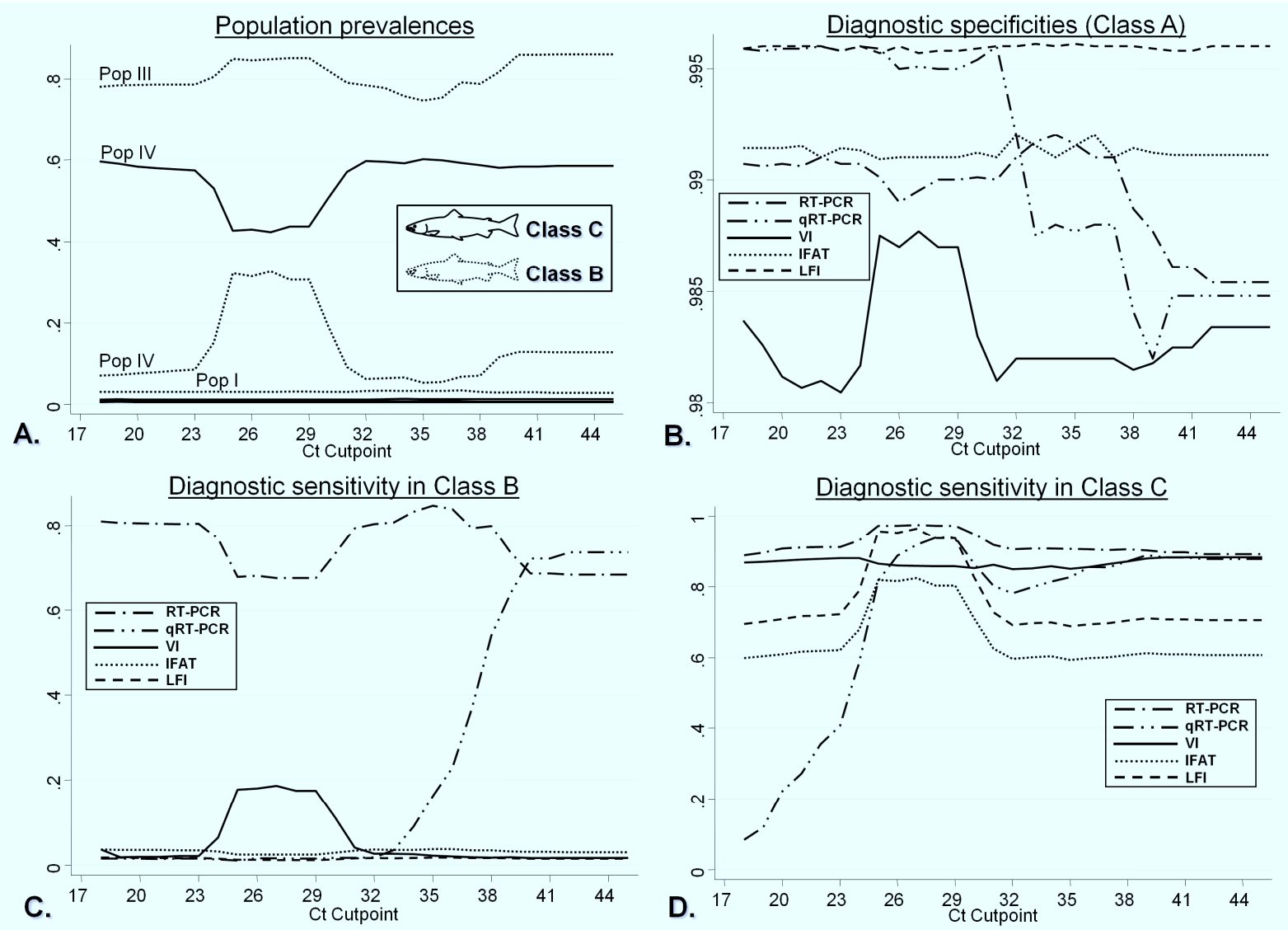


Fig. 4.8. 3-class LCM posterior means of prevalences and test operating characteristics across qRT-PCR cycle threshold (Ct) cutpoints. Ct values ranged from 17.29 to 41.86 and the reaction was stopped at 45 cycles. Ct cutpoint was defined as that point where any sample that yielded a score above the specified value were deemed negative. RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay; DSe: diagnostic sensitivity; DSp: diagnostic specificity; 3 classes of fish (A, B & C) where B & C are assumed infected.

IV) (Fig. 4.8A). Consequently, for VI, the DSp increased (Fig. 4.8B) and the DSe for class B fish increased (Fig. 4.8C). For the other tests, this reclassification resulted in a decreased DSe in class B for RT-PCR (Fig. 4.8C), and in an increased DSe in class C for all 4 tests, especially for ABAs (Fig. 4.8D).

4.4 Discussion

4.4.1 Validity of the model

Addition of a third class and covariance terms increased the number of parameters to estimate in the model and may have compromised the identifiability of the final model. Adapting from Jones et al. (2010), the number of unknown parameters for a conditionally saturated 3 class model with 5 tests and 4 populations would be 101 ($C(2^P-1)+ K(C-1)$). With a maximum number of covariance parameters (i.e. saturarted), the degrees of freedom (df = 124) in the dataset would be still larger and, therefore, the final model was assumed identifiable. This general rule (degrees of freedom greater that number of parameters) may not, however, be adequate to ensure model identifiability (Goodman, 1974; Jones et al., 2010), and Jacobian ranking of the model would be a more appropriate approach to assess identifiability (Goodman, 1974; Jones et al., 2010). Eventually, only 2 covariance parameters were kept to maintain a comfortable margin and assume identifiability (25 model parameters total). Removing one of the five tests (60 degrees of freedom for 21 to 22 parameters) did not impact the model estimations (Table 4.5). However, with only 3 tests in the model, the difference between degrees of freedom (df =

28) and numbers of parameters (17) decreased and the models may have identifiability issues.

By adding a third class, the model and the data fitness were radically improved compared to the conventional 2LCM, regardless of conditional dependence (Table 4.4). DSe estimates in the 2LCM were approximate averages of the estimates of DSe based on the two infection classes in the 3LCM. This difference in DSe between classes B and C supported the suspicion that the probability to test positive varied among ISAV infected salmon. The mixture distributions of infection classes changed across populations, suggesting a variation of the overall DSe (weighted average of class-specific DSe). Therefore, the initial assumption of a single constant DSe was unreasonable. The assumption of DSe within each infected sub-class seemed more appropriate.

The addition of a third class and covariance terms in the model intended to account for conditional dependence between the two NAATs and the two ABAs. Only covariance terms in the class C were significant confirming initial expectations of test dependence according to the agreement tree (Fig. 4.4) and the similarity of techniques used to detect the same analytical target (Gardner et al., 2000). The bimodal distribution of RT-PCR DSe, and to a lesser degree of qRT-PCR DSe, in class C were explained by the difficulty of the model to identify the dependence between NAATs (Fig. 4.5B). The distribution of the covariance estimates also demonstrated bimodality (data not shown). To a much lesser extent, similar features could be visualized from the DSe distributions of the two dependent ABAs. Overall, the model required further adjustments for conditional dependence, albeit Rindskopf & Rindskopf (1986) assumed that the multiple classes LCM would fully account for conditional dependence. The use of random effects

LCM with multiple classes has also been suggested to adjust for conditional dependence (Dendukuri et al., 2009). Conditional dependence could be avoided by breaking each pair of dependent tests. Unfortunately, analyses of 3-test models were unstable and inconclusive (Table 4.5). With results from only one NAAT and one ABA, the data seemed not to support the existence of a third testing pattern resulting in decreased NAAT DSp (Table 4.5). When removing only one test at a time (4-test models), none of the tests seemed to have a particular impact on the identification of the third class, despite the absence of VI, resulting in some variation of tests' DSe in class C (Table 4.5). The influence of VI classification was further confirmed with the instability of estimates for qRT-PCR cutpoint between 25 and 30 cycles when data were weaker (Fig. 4.8). Ultimately, use of multiple latent classes in LCM required data that supported clear and distinct supplementary testing pattern.

In summary, the addition of a third class addressed concerns regarding assumptions of Hui & Walter for LCM and resulted in good data fit. However, other models might fit the data as well (i.e. similar Bayesian *P*-value). As an addendum to statistical consideration, biological considerations should be used to select and interpret suitable models (Dendukuri et al., 2009).

4.4.2 Interpretation of the third class

In LCM, even with 2 classes, the identification of the classes is subjective (Dendukuri et al., 2009). Labelling of latent classes relies on sound biological judgment and often uses estimated prevalences to compare the original clinical information within

the sampled populations. In this instance, class A predominated in Pop I and II (near-zero and low prevalences) and was therefore identified as having non-infected salmon. Only present in Pop IV (cages encountering an outbreak with high prevalence), class C was identified as having infected fish. The class B fish were present at low frequency (~12%) in outbreak cages and at high frequency (~86%) after the outbreak. In addition, class B fish were exclusively detected by NAATs, although not as successfully as class C. Class B was the only class of fish in Pop III and salmon that tested positive to qRT-PCR in this population obtained high range of Ct values (i.e. 33- 42, Fig. 2) corresponding to low load of target. Class B fish remained present after clinical outbreaks and positive for the targeted portion of the 8th RNA segment of the ISAV genome. If the diagnostic objective of the test was detection of disease, the clinical evidence supported class B fish being non-diseased. The diagnostic target is, however, infection with any genotype of ISAV, and two opposing biological interpretations are possible for class B individuals (i.e. infected or non-infected).

One interpretation of the results suggests that class B fish are non-infected salmon that recovered from an infection and carry residual viral RNA fragments (inactive viral particles). In 2LCM, tests are assumed to measure the same underlying latent trait (analytical target) that reflects the true latent infection/disease status (diagnostic target) (Hui & Walter, 1980) (Fig. 4.9A). The presence of more than 2 classes was interpreted as the violation of this assumption with tests that measure different latent variables (Dendukuri at al., 2009). A strong divergence between two analytical targets (low biological correlation) can result in the differentiation of 3 latent classes. The 3 latent classes consist of i) both analytical targets, ii) none of the analytical targets, and iii) either

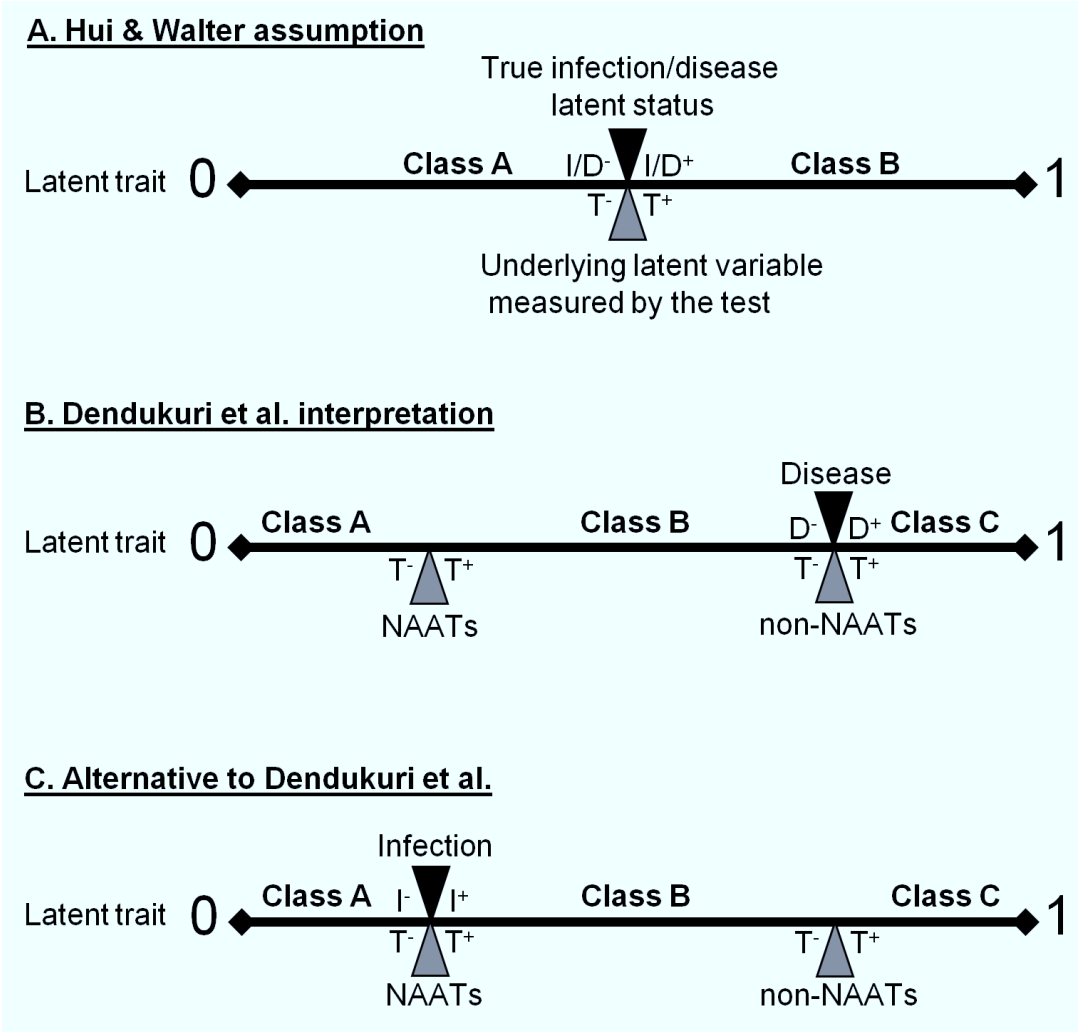


Fig. 4.9. Comparison of the relative correspondence along the latent trait between the true infection status and the latent variable measured by the test(s).

analytical target (i.e. no fourth reciprocal combination). In this study, ABAs had a different analytical target than VI. However, they seemed to be in good agreement compared to the NAATs (Fig. 4.4). Therefore, the 3 assays were assumed to measure a similar latent variable that differed from the one measured by the 2 NAATs. Three classes of samples were indeed observed: salmon that test negative on all tests (no target), salmon that test positive on most or all tests (both targets), and salmon that only test positive on NAATs (only viral RNA target). Tests can measure variables that correspond to the true latent infection status or not. In their study of diagnostic tests for *Chlamydia trachomatis*, Dendukuri et al. (2009) interpreted the non-NAAT assays as measuring a latent variable that matched the true disease status (Fig. 4.9B). Considering that NAATs can detect "infections that are no longer active", the authors assumed that NAATs measured a DNA latent variable that was only a proxy of the true disease status. Therefore, their multiple latent class variable models identified a third class of non-diseased individuals with bacterial DNA. Conversely, in the present study the diagnostic target was infection and not disease. Thus, the first interpretation of class B as non-infected fish assumes that, similar to Dendukuri et al. (2009), the non-NAATs measured the true infection status. This interpretation is supported by the fact that VI is traditionally believed to only detect replicating viral particles and NAATs cannot differentiate between an active and an inactive virus (OIE, 2009b). Devold et al. (2000) could detect ISAV by RT-PCR only in experimentally infected trout up to 135 days post-challenge and explained this observation by the neutralisation (inactivation) of the virus in surviving fish without damaging the integrity of the viral particle. If class B fish are confirmed as non-infected salmon, the probability to test negative (DSp) becomes the

parameter of interest that can be directly computed as the complement of the probability to test positive (1- DSe). Probabilities to test negative in class B were therefore presented in Table 4.3. However, it seemed surprising that the viral particle and associated RNA persist in the organism without being degraded by the host immune system.

The second interpretation considered that class B fish are infected salmon with low numbers of active viral particles and were either a recently infected fish (early stage) or a chronic carrier (recovering from the infection). Contrary to Dendukuri et al. (2009), this interpretation assumes that the true latent infection status is measured by the NAATs (Fig. 4.9C). This interpretation is supported by the fact that low numbers of active virus could only be detected by assays with high analytical sensitivity (ASe), and that the intended diagnostic target included avirulent genotype of ISAV (HPR0) that cannot be detected by VI (Kibenge et al., 2004). First, higher analytical sensitivity and earlier detection of ISAV infection by NAATs compared to ABAs and VI is widely documented and accepted (e.g. Snow et al., 2003; Giray et al., 2005). The limit of detection of qRT-PCR was shown to be 100 times lower than conventional one-tube RT-PCR (Munir & Kibenge, 2004). In addition, qRT-PCR has been shown to detect virus that replicate (i.e. active particles) without production of CPE in cell cultures (Kibenge et al., 2004).

Therefore, NAATs seemed to be more sensitive to detect early infected fish that carry a low number of active particles. Furthermore, experimental challenge revealed that the level of infection decreased at 15 days post-infection due to of the host immune response (Mikalsen et al., 2001). Although ISA is often acute, chronic forms of the infection have been proposed based on antibody profiles (Kibenge et al., 2002). Fish that recovered from infection can still transmit ISAV by cohabitation with healthy fish despite

testing negative to VI (Kibenge et al., 2004). Therefore, NAATs may detect low level carrier stages after clinical outbreaks that were not detected by other tests. In these fish, however, the concentration of viral particles may be close to the limit of detection of the NAATs, resulting in detection inconsistencies (lower DSe) (McAllister et al., 2003) and decreased agreement between the 2 assays. Furthermore, the analytical specificity (ASp) differed amongst tests: the avirulent genotype HPR0 was detected by RT-PCR but not by VI (Kibenge et al., 2004). In our study, class B salmon may have been HPR0 infected fish since this isolate is present in the region, albeit this was not confirmed or verified in this study. Although there is no evidence that ABAs are not able to detect HPR0, low level of replication of this virus type may explain the rare detection by these assays (Gustafson et al., 2008). However, the outbreak cages where most class B fish were sampled were originally diagnosed with virulent types (HPR4 and HPR2). Also, the large proportion of class B fish in Pop III (~ 86%) did not suit previous descriptions of HPR0 infection in the region (i.e. low proportion of fish infected) (Dr. Mike Beattie, pers. com.). In general, diagnostic tests dichotomize the measurement of an underlying biological trait either using a cutpoint or simply due to their limit of detection (Brenner & Gefeller, 1997). As a result, this second interpretation suggested that NAATs measured the true latent infection status and the non-NAATS, limited by their ASe and ASp, dichotomized the infected class into 2 sub-classes of low and high infected salmon (Fig. 4.9C).

The lack of detailed information on ISAV infection dynamics and a weak understanding of the correlation between the analytical target and the intended diagnostic target compromised the interpretation of this supplementary class. For instance,

transmission studies under controlled conditions would determine the infectiveness of fish that are only detected by NAATs at different stages of the infection (i.e. pre- and post-clinical manifestation). Class B fish could also be a mix of low-infected and convalescent fish. In addition, at least a portion of the samples that tested positive in Pop III (class B fish) should be sequenced to determine if some were HPR0 infected. Overall, the intended purpose and associated diagnostic target of a test should be clearly defined before evaluation, and a good understanding of its correlation with the analytical target is critical for the subsequent interpretation of multiple classes.

4.4.3 Past evaluations and interpretation

A direct comparison of test estimates of this study with previous ISAV test evaluations would be inappropriate since some tests did not use identical protocol (e.g. RT-PCR or VI) and DSe and DSp are considered population-dependent (Greiner & Gardner, 2000). Furthermore, even if DSp estimates maybe be compared, DSe were split between class B and C in this study and no longer correspond to the definition. For instance, IFAT was the only standard assay used by all previous studies (McClure et al., 2005; Nérette et al., 2005; Gustafson et al., 2008; Nérette et al., 2008a). Although DSp estimates were fairly similar, DSe differed substantively, except when the estimation accounted for dependence (Nérette et al., 2008a). This last comparison may be of little value since Nérette et al. (2008a) did not differentiate three classes of fish. However, the 2LCM analysis conducted by Nérette et al. (2008a) considered non-infected samples that only tested positive to RT-PCR and estimated then a lower DSp for this assay. Therefore,

the estimated DSe corresponded to samples that test positive to several assays and is comparable to the DSe in class C. Comparisons of other tests would be futile given the discrepancies in test protocols and the complexity of the interpretation.

In this evaluation, the intended purpose was to use ISAV detection tests in surveillance programs to demonstrate population freedom from infection (i.e. cage, site, bay, region, country). In this instance, the tested population is assumed free and, if detected, the infection is assumed to exist at an early stage of the infection. For domestic surveillance, the parameter of interest is the positive predictive value of a positive test result (PPV). When the geographical region is not free of disease, the surveillance objectives change focusing on early detection of the infection to implement control measures (e.g. depopulation). In this instance, the parameter of interest is negative predictive value of a negative test result (NPV). Predictive values depend on the test operating characteristics (DSe/DSp) and the assumed prevalence in the sampled population (Dohoo et al., 2009). However, the applications of DSe and DSp traditionally involved two classes of animals, and an adjustment from the conventional approach would be required to apply three classes of information. A good understanding of the infection dynamics within the populations is then necessary to appreciate the proportion of the 3 classes of fish at different stages of a disease event and predict predictive values. In addition, the interpretation of test results relies on the context of its utilisation.

In Atlantic Canada, the control measures require that any cage declared infected should be culled within 7 days (DAA, 2007). Therefore, if class B fish are confirmed to be recovering fish, the probability of sampling these fish in this region is almost nil. If class B fish are, however, confirmed as low-infected fish, the proportion of low- and

high-infected fish at the initial stage of the infection are critical for diagnostic test selection in this region. By comparison, in Norway, when a cage is declared infected, then the whole farm would be depopulated but within a flexible period of 80 days (Norwegian Food Safety Authority, 2007). Therefore, it appears critical that detailed descriptions of the ISAV infection dynamics and associated clinical information be investigated at the fish and population level to enhance the interpretation and application of multiple class models. Nonetheless, the computation of predictive values assumes that sampling of fish is representative of the targeted population (i.e. random sample). In practice, however, dead or moribund salmon are conveniently/purposefully targeted (DAA, 2007). As a consequence, the proportions of diseased fish (prevalence) may differ from the population of live fish.

4.5 Conclusion

This study supports the use of multiple classes in LCM of diagnostic tests for ISAV in salmon when the assumption of constant DSe (or DSp) is inappropriate and when data clearly show distinct testing patterns. More than two classes should also be considered when tests target very different biological analytes and when the analytical target is weakly associated with the intended diagnostic target. Clearly defined diagnostic purposes and biological understanding are essential to subsequently identify the different classes needed for the LCM. Even with additional classes, the assumption of conditional independence should be verified and, if necessary, adjusted. Identifiability might restrict the use of multiple classes for models when a limited number of tests and populations are

available. Addition of informative priors can provide an alternative, but should be used with caution. With multiple classes, the evaluation and interpretation of diagnostic tests, not only for ISAV in salmon but for other pathogens and in other species, is evolving and requires further understanding of the infection/disease dynamics at the animal and population levels for relevant applications to surveillance situations.

4.6 References

Alberg, A.J., Park, J.W., Hager, B.W., Brock, M.V., Diener-West, M., 2004. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J. Gen. Intern. Med. 19, 460-465.

Begg, C.B., 1987. Biases in the assessment of diagnostic tests. Stat. Med. 6, 411-423.

Branscum, A.J., Gardner, I.A., Johnson, W.O., 2005. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling**.** Prev. Vet. Med. 68, 145-163.

Brenner, H., 1996. How independent are multiple "independent" diagnostic classifications? Stat. Med. 15, 1377-1386.

Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat. Med. 16, 981-991.

Brooks, S.P., Gelman, A., 1998. Alternative methods for monitoring convergence of iterative simulations. J. Comp. Graph. Stat. 7, 434-455.

DAA, 2007. Infectious salmon anemia (ISA) management and control program. Department of Agriculture and Aquaculture. New Brunswick, Canada. 32pp.

Dendukuri, N., Joseph, L., 2001. Bayesian Approaches to Modeling the Conditional Dependence Between Multiple Diagnostic Tests. Biometrics 57:158-167.

Dendukuri, N., Hadgu, A., Wang, L., 2009. Modeling conditional dependence between diagnostic tests: A multiple latent variable model. Stat. Med. 28:441-461.

Devold, M., Krossøy, B., Aspehaug, V., Nylund, A., 2000. Use of RT-PCR for diagnosis of infectious salmon anaemia virus (ISAV) in carrier sea trout *Salmo trutta* after experimental infection. Dis. Aquat. Organ. 40, 9-18.

Dohoo, I., Martin, W., Stryhn, H. (Eds.), 2009. Veterinary Epidemiologic Research. 2nd ed., AVC Inc., Charlottetown, Canada.

Espeland, M.A, Handelman, S.L., 1989. Using latent class models to characterize and assess relative error in discrete measurements. Biometrics 45, 587-599.

Falk, K., Dannevig, B.H., 1995. Demonstration of infectious salmon anaemia (ISA) viral antigens in cell cultures and tissue sections. Vet. Res. 26, 499–504.

Falk, K., Namork, E., Dannevig, B.H., 1998. Characterization and applications of a monoclonal antibody against infectious salmon anaemia virus. Dis. Aquat. Org. 34, 77–85.

Formann, A.K., 1994. Measurement errors in caries diagnosis: some further latent class models. Biometrics 50, 865-871.

Gardner, I., Stryhn, H., Lind, P., Collins, M., 2000. Conditional dependence between tests affects the diagnosis and surveillance of animal disease. Prev. Vet. Med. 45, 107-122.

Gelman, A., 1996. Inference and monitoring convergence, Chapter 8. In: Gilks, W., Richardson, S., Spiegelhalter, D. (Eds.), Markov Chain Monte Carlo in practice. Chapman et Hall, London, UK, pp. 131-140.

Georgiadis, M.P., Johnson, W.O., Gardner, I.A., Singh, R., 2003. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. Appl. Stat. 52, 63-76.

Giray, C., Opitz, H.M., MacLean, S., Bouchard, D., 2005. Comparison of lethal versus non-lethal sample sources for the detection of infectious salmon anemia virus (ISAV). Dis. Aquat. Organ. 23, 181-185.

Godoy, M.G., Aedo, A., Kibenge, M.J., Groman, D.B., Yason, C.V., Grothusen, H., Lisperguer, A., Calbucura, M., Avendaño, F., Imilán, M., Jarpa, M., Kibenge, F.S., 2008. First detection, isolation and molecular characterization of infectious salmon anaemia virus associated with clinical disease in farmed Atlantic salmon (Salmo salar) in Chile. BMC Vet. Res. 4, 28.

Goodman, L.A., 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61, 215–231.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 42, 2-22.

Gustafson, L., Ellis, S., Bouchard, D., Robinson, T., Marenghi, F., Warg, J., Giray, C., 2008. Estimating diagnostic test accuracy for infectious salmon anaemia virus in Maine, USA. J. Fish Dis. 31, 117-125.

Gustafson, L., Ellis, S., Merrill, P., Robinson, T., MacPhee, D., 2005. Mortality rates predict apparent prevalence of infectious salmon anaemia (ISA) at two infected Atlantic salmon farms in Maine. Bull. Eur. Ass. Fish Pathol. 25, 212-220.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41, 95-98.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.

Hui, S.L, Zhou, X.H., 1998. Evaluation of diagnostic tests without gold standards. Stat. Met. Med. Res. 7, 354-370.

ISO International Standard 5725-1, 1994. Accuracy (trueness and precision) of measurement methods and results. Part 1: General principles and definition. International Organisation for Standardisation (ISO), ISO Central Secretariat, 1 rue de Varembé, Case Postale 56, CH - 1211, Geneva 20, Switzerland.

Johnson, W.O., Gardner, I.A., Metoyer, C.N., Branscum, A.J., 2009. On the interpretation of the test sensitivity in the two-test two-population problem: Assumptions matter. Prev. Vet. Med. 91, 116-121.

Jones, G., Johnson, W.O., Hanson, T.E., Christensen, R., 2009. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics. [Epub ahead of print]

Kibenge, F.S., Munir, K., Kibenge, M.J., Joseph, T., Moneke, E., 2004. Infectious salmon anemia virus: causative agent, pathogenesis and immunity. Anim. Health Res. Rev. 5, 65-78.

McAllister, P.E., Densmore, C.L., Barbash, P.A., 2003. Infectious salmon anaemia virus: injection challenge and waterborne transmission monitored by hematology and polymerase chain reaction assay. In: 28th Annual Eastern Fish HealthWorkshop.

McClure, C.A., Hammell, K.L., Dohoo, I.R., Nerette, P., Hawkins, L.J., 2004. Assessment of infectious salmon anaemia virus prevalence for different groups of farmed Atlantic salmon, Salmo salar L., in New Brunswick. J. Fish. Dis. 27, 375-383.

McClure, C.A., Hammell, K.L., Stryhn, H., Dohoo, I.R., Hawkins, L.J., 2005. Application of surveillance data in evaluation of diagnostic tests for infectious salmon anemia. Dis. Aquat. Org. 63, 119-127.

Mikalsen, A.B., Teig, A., Helleman, A.L., Mjaaland, S., Rimstad, E., 2001. Detection of infectious salmon anaemia virus (ISAV) by RT-PCR after cohabitant exposure in Atlantic salmon *Salmo salar*. Dis. Aquat. Org. 47, 175–181.

Mintiens, K., Toft, N., Lewis, F., Verloo, D., Georgiadis, M., Johnson, W., Gardner, I., Wright, P., Gunn, G., Greiner, M., . Guidelines on the use of no-gold standard methods for estimating diagnostic characteristics of microbiological and serological assays. OIE Sci. Tech. Rev. (submitted).

More, S.J., Cameron, A.R., Greiner, M., Clifton-Hadley, R.S., Rodeia, S.C., Bakker, D., Salman, M.D., Sharp, J.M., De Massis, F., Aranaz, A., Boniotti, M.B., Gaffuri, A., Have, P., Verloo, D., Woodford, M., Wierup, M., 2009. Defining output-based standards to achieve and maintain tuberculosis freedom in farmed deer, with reference to member states of the European Union. Prev. Vet. Med. 90, 254-67.

Munir, K., Kibenge, F.S.B., 2004. Detection of infectious salmon anaemia virus by real-time RT-PCR. J. Virological Methods 117, 37–47.

Neath, A.A., Samaniego, E.J., 1997. On the efficacy of Bayesian inference for nonidentifiable models. Am. Stat. 51: 225-232.

Nérette, P., Dohoo, I., Hammell, L., 2005. Estimation of specificity and sensitivity of three diagnostic tests for infectious salmon anaemia virus in the absence of a gold standard.. J. Fish Dis. 28, 89-99.

Nérette, P., Stryhn, H., Dohoo, I., Hammell, L., 2008a. Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic tests. Prev. Vet. Med. 85, 207-225.

Nérette, P., Hammell, L., Dohoo, I., Gardner I., 2008b. Evaluation of testing strategies for infectious salmon anaemia and implications for surveillance and control programs. Aquaculture 280, 53-59.

Norwegian Food Safety Authority, 2007. Contingency plan for control of infectious salmon anaemia (ISA) in Norway. (accessed 23 May, 2007), http://www.mattilsynet.no/mattilsynet/multimedia/archive/00025/ILA_contingensy_plan_25394a.pdf.

Office International des Epizooties, 2009a. OIE Aquatic Animal Health Code. 12th Edition. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 99-104.

Office International des Epizooties, 2009b. Manual of Diagnostic Tests for Aquatic Animals 2009. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 10-30.

Qu, Y., Tan, M., Kutner, M.K., 1996. Random effects models for evaluating accuracy of diagnostic tests. Biometrics 52, 797-810.

Ransohoff, D.F., Feinstein, A.R., 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N. Engl. J. Med. 17, 926-930.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009) *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Rindskopf, D., Rindskopf, W., 1986. The value of latent class analysis in medical diagnosis. Stat. Med. 5, 21-28.

Rolland, J.B., Bouchard, D., Coll, J., Winton, J.R., 2005. Combined use of ASK and SHK-1 cell lines to enhance the detection of infectious salmon anemia virus. J. Vet. Diagn. Invest. 17, 151-157.

Snow, M., Raynard, R.S., Murray, A.G., Bruno, D.W., King, J.A., Grant, R., Bricknell, I.R., Bain, N., Gregory, A., 2003. An evaluation of current diagnostic tests for the detection of infectious salmon anaemia virus (ISAV) following experimental water-borne infection of Atlantic salmon, Salmo salar L. J. Fish. Dis. 26, 135-45.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A., 2002. Bayesian measures of model complexity and fit. J.R. Stat. Soc. B 64, 583–639.

Spiegelhalter, D.J., Thomas, A., Best, N.G., 2003. WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596-1599.

Thurmond, M.C., Johnson, W.O., 2004. Effect of multiple sampling on diagnostic sensitivity. J. Vet. Diagn. Invest. 16, 233-236.

Toft, N., Jorgensen, E., Hojsgaard, S., 2005. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev. Vet. Med. 68, 19–33.

Toft, N., Innocent, G., Gettingby, G., Reid, S., 2007. Assessing the convergence of Markov Chain Monte Carlo methods: An example from evaluation of diagnostic tests in absence of a gold standard. Prev. Vet. Med. 79, 244-256.Agresti, A., 2002. Categorical data analysis. New York: Wiley.

Torrance-Rynard, V.L., Walter, S.D., 1997. Effects of dependent errors in the assessment of diagnostic test performance. Stat. Med. 16, 2157-2175.

Vacek, P.M., 1985. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 41, 959-968.

Wilson, I.G., 1997. Inhibition and facilitation of nucleic acid amplification. Appl. Environ. Microbiol. 63, 3741-3751.

Yang, I., Becker, M.P., 1997. Latent variable modeling of diagnostic accuracy. Biometrics 53, 948-958.

Yerushalmy, J., 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Publ. Health Rep. 62, 1432-1449.

Chapter V: SELECTION OF A CUTPOINT VALUE FOR REAL-TIME PCR RESULTS TO FIT A DIAGNOSTIC PURPOSE: ANALYTICAL AND EPIDEMIOLOGICAL APPROACHES

Abstract

Diagnostic laboratories frequently select an arbitrary cutpoint value for real-time amplification assays above which an obtained cycle threshold (Ct) value is deemed false. High Ct values are interpreted as amplification or fluorescence artifacts, or cross-contaminations. However, the Ct cutpoint should be chosen with rational justification. This study reviewed analytical criteria to select cutpoints during the development of the assay, including fluorescence threshold, reaction end cycle, limit of detection, artifact investigation. The degree of variation of amplification efficacy within and between laboratories may result in cutpoint changes across runs requiring standardization procedures for the Ct cutpoints. Selection strategies were further reviewed based on epidemiological parameters considering the probability or the cost of a false test result based on a specified cutpoint. Depending on the intended purpose of the test, cutpoints can be selected graphically to maximize the probability of either true positive or true negative using the Two-Graph Receiver Operating Characteristics (TG-ROC). Diagnostic sensitivity and specificity may vary with the tested population, therefore, the estimated TG-ROC curve is population-dependent and should be validated for a specified purpose. Although the selection of a cutpoint based on misclassification cost depends on infection prevalence, the selection based on predictive values does not.

5.1 Introduction

In the past decade, polymerase chain reaction (PCR) tests have become a standard method for the detection of a wide range of pathogens and biomarkers in routine veterinary diagnostic testing. Lately, laboratories are progressively shifting from conventional PCR to a method referred to as kinetic, quantitative or real-time PCR (qPCR), which allows for quantification based on monitoring of the progression of the amplification during the cyclic reaction. Compared to the standard PCR, this method provides substantial benefits to laboratorians, including: (i) reduced time of analysis, (ii) increased analytical sensitivity (ASe), (iii) increased reproducibility, and (iv) decreased cross-contamination (Mackay et al., 2002). Although the cost of equipment and reagents restricted the utilization of this method in the past (Mackay et al., 2002), qPCR is now more affordable making it a key technology used by many diagnostic laboratories today.

Real-time outcome

The resulting outcome of the qPCR is continuous, while that provided by conventional PCR is binary (i.e. absence or presence of an expected gel band). Referred to as the cycle threshold (Ct) or crossing point (CP) value, the qPCR outcome reflects the cycle value at which the fluorescence signal exceeds a defined background threshold (Fig. 5.1). Real-time PCR results differ from continuous outcomes of other diagnostic assays (e.g. ELISA) in that negative specimens do not yield Ct values since the fluorescent signal stays below the specified threshold. The distributions of Ct values are generally non-normal, heteroscedastic and truncated (Burns & Valdivia, 2008). A more

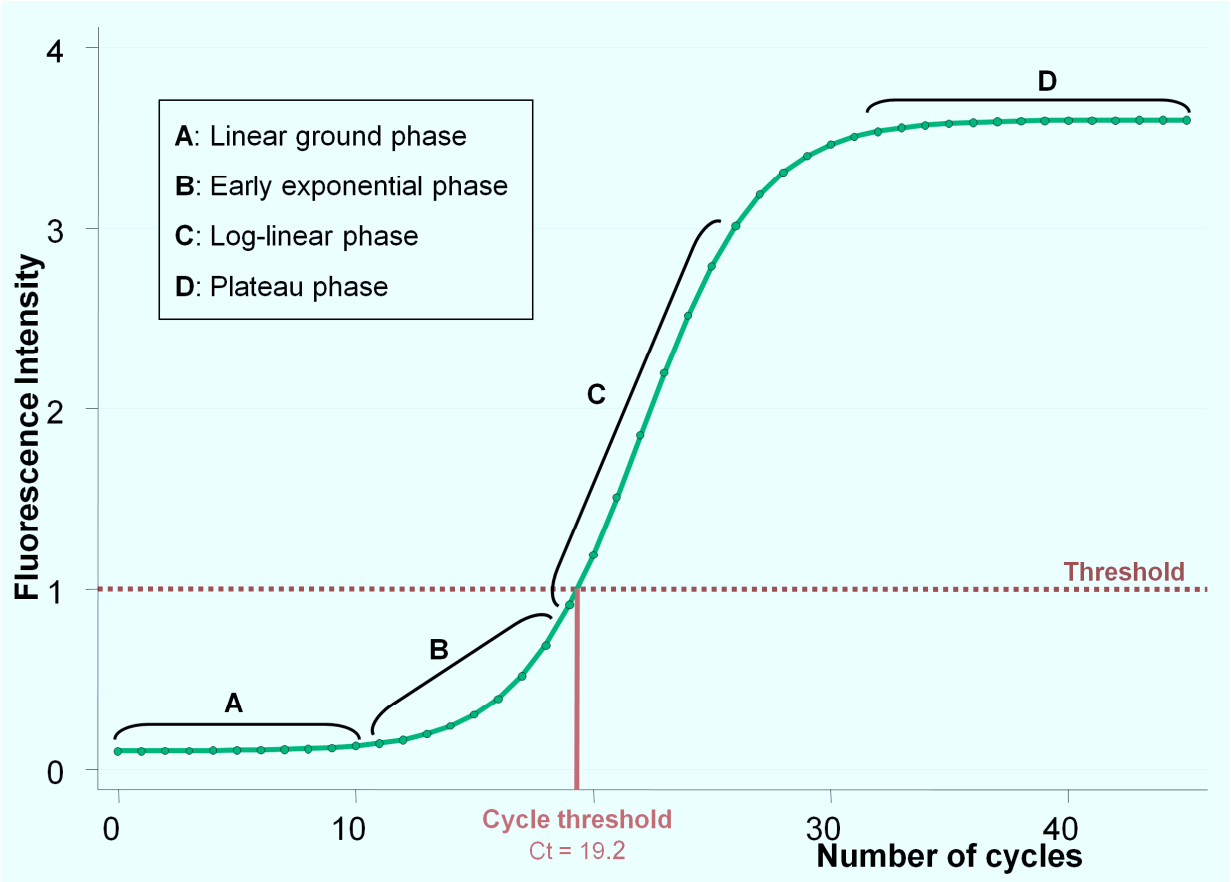


Fig. 5.1. Sigmoid shaped profile of fluorescence accumulation across cycles during a real-time amplification. The amplification curve experiences four different phases: linear ground (A); early exponential (B); log-linear (C); and plateau (D). A fluorescence threshold is set where the amplification curve is beginning the log-linear phase. The threshold level is determined using either complex algorithms or arbitrarily. The intersection between the threshold and the curve is the cycle threshold (Ct) value.

appropriate description of the qPCR outcome would, therefore, be *hemi-continuous*. In addition, the quantitative Ct value is inversely proportional to the (log) concentration of nucleic acids detected (i.e. high Ct value reflects a low target concentration and *vice versa*).

Traditionally, regardless of the value intensity, when any Ct value is produced, the specimen tested is deemed to be positive. Recently, however, there is an increasing tendency among laboratory operators to consider as negative (i.e. false positive) any Ct values above a subjective cutpoint value. It is assumed that the high Ct value is generated by either the degradation of the probe-based fluorophore, cross-contamination, or non-specific amplification of background nucleic acids (Burns & Valdivia, 2008). A sample with a Ct value greater than the cutpoint might, therefore, would be classified and reported to the end user as a negative result with no way to distinguish it from results that did not produce a Ct value.

The objective of this work was to review published justifications based on biologically sound rationale for choosing cutpoint values in real-time PCR. Approaches for the selection of Ct cutpoint can be considered at the bench level (analytical approach) and at the population level (epidemiological approach) (Fig. 5.2). The application of the epidemiological principles was illustrated with the example of a quantitative reverse-transcriptase PCR (qRT-PCR) used to detect the presence of infectious salmon anaemia virus (ISAV) in farmed Atlantic salmon in Canada.

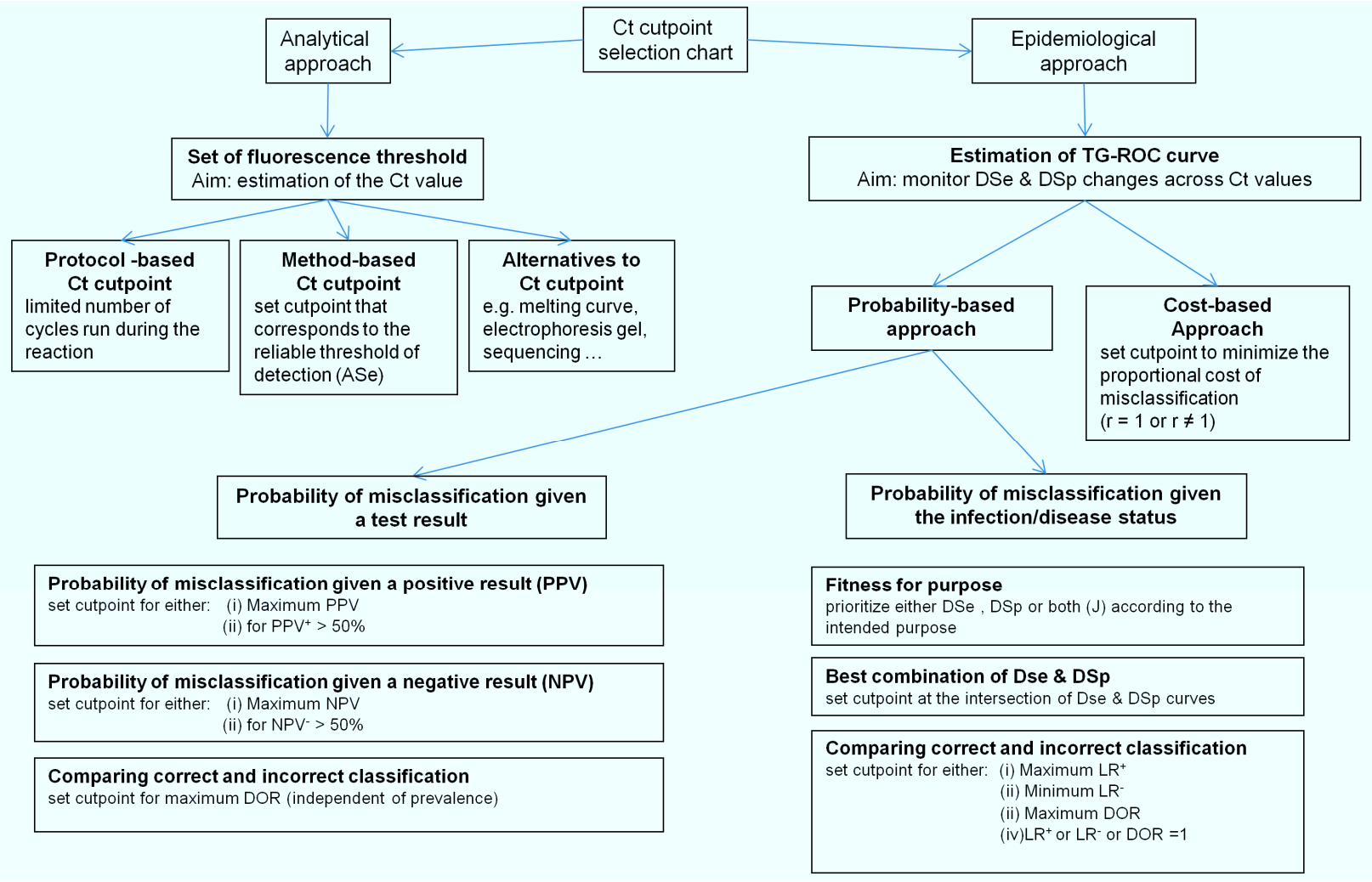


Fig. 5.2. Decision tree for selecting a cutpoint for real-time amplification assays.
DSp: diagnostic specificity; DSe: diagnostic sensitivity; J: Youden index (DSe + DSp -1); DOR: diagnostic odds ratio; LR^+: likelihood ratio of a positive test; LR^-: likelihood ratio of a negative test.

5.2 Empirical justifications of cutpoints for real-time PCR assay

5.2.1 Justifications and selection of cutpoint at the "bench" level (analytical)

Analytical criteria refer to technical parameters of the assay evaluated during the bench development, optimization and standardization of the method.

5.2.1.1 Fluorescence signal threshold

The first consideration in the development of a qPCR protocol is the selection of a threshold level for the fluorescence signal. This limit is either selected by applying complex algorithms (Rebrinov & Trofimov, 2006) or visually by the operator. The approach to select the threshold is often directed by the equipment and/or the software associated with the thermocyclor. A simple approach is to set the threshold at several standard deviations above the baseline mean (Wong & Medrano, 2005) (Fig. 5.1), as it is done with other continuous assays (e.g. ELISA) (Ambruster et al., 1994). However, the intention in qPCR is, however, to ensure that the log-linear phase of the amplification is reached (Ruttledge, 2004) (Fig. 5.1), and not that a targeted proportion of the non-infected/disease population is selected (Sunderman, 1975). Although the fluorescence signal threshold inherently influences the position of a Ct value, the operator primarily focuses on the discrimination among the Ct values themselves.

5.2.1.2 Limited number of cycle (amplification efficacy)

Technically, qPCR already uses a cutpoint for Ct values by maximizing the number of amplification cycles. Hypothetical increases of the fluorescence signal above the threshold after the last reaction cycle are not detected and considered negative. In general, the number of run cycles is set on the assumption that if a single copy of the target is present in the tested specimen, amplicons should be generated in sufficient quantity to be detected before the last cycle. Usually, qPCR protocols are set up to 40 cycles which yields, in theory, a trillion amplicons from a single target molecule (Table 5.1), assuming that the number of copies doubles at each reaction (i.e. amplification efficacy (E) = 1). Amplification efficacy is, however, rarely equal to unity and the amplification may decrease as the reaction progresses due to the decline of reaction reagents (Mehra & Hu, 2005). Therefore, the numbers of generated amplicons can differ greatly (Table 5.1). Even if the qPCR is already limited by the number of run cycles, the operator may still select a Ct cutpoint before the last cycle. One technical justification to select a Ct cutpoint earlier during the reaction is to consider a sample negative when it is below the corresponding reliable limit of detection of the assay.

5.2.1.3 Cutpoint as the limit of detection

During the development of an assay, the linear operating range of the method must be determined, including the estimation of the lower and upper limits of detection (OIE, 2009). Referred to as analytical sensitivity (ASe), the lower limit of detection is

Table 5.1

Correspondence between number of amplicons generated (X_n) and the number of cycles (n), according to the amplification efficacy (E). This table was generated using the formula: $X_n = X_0 (1+E)^n$ (adapted from Rebrikov & Trofimov, 2006). X_0 refers to the initial number of target copies at cycle 0 (single copy here).

Number of cycles	X_n (E = 1)	X_n (E = .9)
10	1,024	613
20	1,048,576	375,899
30	1,073,741,824	230,466,617
40	1,099,511,627,776	141,300,610,453

defined as the ability of the test to detect a minimum concentration of analyte with a known certainty (OIE, 2009). According to the OIE (2009), the ASe is determined as the end-point dilution at which 50% of the tested samples are positive. For PCR, the limit of "*exactly 50%*" positive test results is a recent modification from the previous international requirement of "*at least 95%*" (Burns & Valdivia, 2008). Like most binary outcome diagnostic tests, qPCR is based on the dichotomization of individuals according to a measured continuous underlying trait (i.e. target concentration) (Brenner & Gefeller, 1997). An analytical cutpoint could therefore be justified by selecting a Ct value corresponding to the lower limit of detection. Any Ct value above this defined limit (i.e. lower amount of target) would not be considered reliable.

Burns & Valdivia (2008) suggested two approaches to estimate the ASe. The first, referred to as "Experimental" ASe (ASe_{exp}), determines the last investigated serial dilution where at least 95% of the samples tested positive (Burns & Valdivia, 2008). This approach is less precise since the exact concentration associated with 95% detection is not estimated; however, the design and the estimation are more forgiving and straightforward. In Burns & Valdivia (2008), the number of replicates per concentration did not influence significantly the estimation of ASe_{exp}. For the sake of simplicity, we suggest adapting the ASe_{exp} estimation approach under the current OIE requirements of 50% of positive test results associated to a certain degree of risk (type I error, α). This approach would lower the necessary number of samples tested for each dilution. For instance, the sample size calculation can be based on the computation to demonstrate freedom of disease (Dohoo et al., 2009). To demonstrate, at a specified concentration, that the assay detects at least 50% of samples (i.e. proportion of negative test results

lower than 50%) with a risk level set at $\alpha = 5\%$ (95% confidence), a minimum of 5 replicates per dilution would be necessary. The ASe_{exp} will then be estimated as the last serial concentration that yields positives on all 5 replicates. For 99% level of confidence, 7 replicates would be necessary. Furthermore, it would be of value to specify the dilution factor used in the serial dilution. The lower the dilution factor, the more refined is the estimation (i.e. smaller gap between serial dilutions).To fit the log scale, dilution factors of 10 are frequently used albeit Burns & Valdivia (2008) used a factor of 2 to increase the precision of estimation.

The second approach, referred to as "Theoretical" ASe (ASe_{theo}), is estimated using computer-based regression modelling where the exact theoretical dilution of the 95% positives is estimated (Burns & Valdivia, 2008). Although more complex and exacting, this method is more sensitive to the number of replicates per concentration. In addition, caution should be taken when using this approach since the regression computation requires certain assumptions that may not be reasonable with the atypical nature of Ct values (e.g. heteroscedasticity).

Once the ASe has been estimated, the Ct cutpoint is selected as the Ct value corresponding to the estimated ASe. However, the efficacy of amplification can differ greatly across reactions. Further standardization can be achieved by either estimating the Ct value corresponding to the ASe using a standard curve (absolute quantification approach), or by normalization of the Ct cutpoint using the Ct value of a reference gene co-amplified in multiplex or parallel (relative quantification approach).

In conclusion, the selection of a Ct cutpoint can be based on the linear operating range of the assay. Specimens with pathogen concentrations lower than the ASe are,

however, considered negative when they can still be detected (i.e. with a probability lower than 50%). An alternative approach is to develop verification techniques to investigate cross-contaminations and potential amplification artifacts.

5.2.1.4 Investigation of artifactual results

Amplicon artifacts are spurious products or primer dimers that are usually observed toward the end of the reaction. According to the fluorophore chemistry employed, further procedures are routinely used to discriminate fluorescence signal artifacts. For instance, in SYBR green-based assays, investigation methods include melting curve, gel electrophoresis and sequencing. Comparatively, probe-based assays (e.g. Taqman), which are based on the detection of a specific sequence within the amplicon, limit potential non-specific fluorescence signal.

Alternatively, cross-contamination from positive controls to test samples is the other common type of false positive result. In this situation, the intended strategy is to use positive controls associated with an original marker that allows tracking of potential contamination afterward. Although the concept stays the same, a wide range of marking strategies and marker detection techniques exist. For instance, the design of a plasmid-based positive control containing a unique target site for a probe is one such approach (Snow et al., 2009). The positive control is detected by the first probe specific to the target, and also by a second probe (different fluorophore) targeting a unique sequence. Both probes are added to test samples, and positive results are detected by the first probe.

The accidental introduction of the positive control plasmid to the sample will induce the second probe to also produce a signal.

5.2.2 Justifications and selection at the "population" level (diagnostic performance)

Factors that influence diagnostic misclassification of a test include parameters at the bench level and errors made in the "field" during sample collection and storage, which can greatly influence the interpretation and reporting of test results (OIE, 2009). An epidemiological approach that accounts for all these factors provides a pragmatic evaluation of a detection method and assists authorities in the decision-making process. Epidemiology-based justifications for the selection of a Ct cutpoint are twofold: i) to minimize the probability of misclassification; and ii) to minimize the cost associated with misclassification. The parameters of interest that quantify the epidemiological operating performances of an assay, and the justifications for Ct selection, will be defined, and the graphical tool used to monitor them will be described.

5.2.2.1 Test operating characteristics

5.2.2.1.1 Diagnostic accuracy

The accuracy of a diagnostic test is traditionally reported separately within infected/diseased (D+) and non-infected/non-diseased (D-) individuals. Diagnostic sensitivity (DSe) refers to the probability of a specimen testing positive given that the

sampled individual is D+, and diagnostic specificity (DSp) refers to the probability of a specimen testing negative given that the sampled individual is D- (Yerushalmy, 1947).

Overall assay performances can also be expressed using single parameters. The Youden index (J) expresses the average of "successes" of the test (difference between proportions of correctly classified and incorrectly classified) within the D+ group and within the D- group (Youden, 1950), and was suggested as a potential criterion to select cutpoints (Greiner et al., 1995). The estimation of J can be directly computed from DSe and DSp as follows:

$$J = DSe + DSp - 1 \qquad (1)$$

The test discrimination is deemed positive (i.e. test useful) when J is above 0.

The test efficiency (Ef) (proportion of correctly classified samples) is another single parameter expressing accuracy. Computed as Ef = Pr * DSe + (1-Pr) * DSp (Greiner et al. 2005), this parameter is, however, of less value since it depends on Pr, the prevalence of D+ (Alberg et al., 2004).

The DSe and Dsp are estimated during test evaluation studies that imply that the health status of tested specimens is known with certainty. In practice, the status is unknown and the test result is the only information available. As a result, a test user seeks the probability of D+ given a particular test result. This corresponds to estimating the positive predictive value of a positive test result (PPV), which refers to the probability that an individual is D+ given that its specimen tested positive. Alternatively, the negative predictive value (NPV) corresponds to the probability that an individual is D-

given that its specimen tested negative (Vecchio, 1966). Predictive values depend on DSe, DSp, and on Pr in the tested population, as shown in the following formulae:

$$PPV = [DSe * Pr] / [DSe * P + (1 – DSp) * (1 – Pr)] \quad (2)$$

$$NPV = [DSp * (1 – Pr)] / [(1 – Dse) * Pr + DSp * (1 – Pr)] \quad (3)$$

Diagnostic accuracy can also be expressed as likelihood ratios (LR). In the situation of a binary test result (i.e. a cutpoint specific LR), the likelihood ratio for a positive test result (LR+) reflects how much more likely D+ individuals are to test positive compared to D- individuals. The likelihood ratio for a negative test result (LR-) reflects how much less likely D+ individuals are to test negative compared to D- individuals (Dohoo et al., 2009). Likelihood ratios can be defined for specific cutpoints associated with a different set of DSe/DSp and are estimated using the following formulae:

$$LR+ = DSe / (1 - DSp) \quad (4)$$

$$LR- = (1-DSe) / DSp \quad (5)$$

The last parameter of interest is diagnostic odds ratio (DOR) and it is another single measure of diagnostic accuracy that reflects the ratio of the odds of infection/disease in positive test results over the odds of infection/disease in negative test results (Glas et al., 2003). DOR is therefore estimated using either LR+/LR-, DSe/DSp, or PPV/NPV:

$$DOR = LR^{+}/LR^{-} = [DSe/(1-DSe)]/[(1-DSp)/DSp] = [PPV/(1-PPV)]/[(1-NPV))/NPV] \quad (6)$$

As the DOR increases above 1, the test becomes more useful. If DOR is between 0 and 1, tested individuals have a greater chance of being misclassified than correctly classified (negative discrimination).

5.2.2.1.2 Two-graph receiver operating characteristic (TG-ROC) plot

Derived from the receiver operating characteristic (ROC) analysis, the two-graph receiving operating characteristic (TG-ROC) plot was first developed by Greiner et al. (1995) to graph the variation of DSe and DSp of a continuous outcome assay across a range of cutpoints. Assuming that the cutpoint value is an independent variable, TG-ROC identifies intermediate test performances and facilitates the graphical selection of cutpoint values. For conventional continuous outcome assays that directly quantify a target, the curve for DSe monotonically decreases toward 0% with an increase of the cutpoint, while the DSp curve increases toward 100%. Conversely, Ct values are inversely proportional to the (log) concentration of the target. Hence, DSe increases toward 100% with increasing Ct values while DSp decreases (Fig. 5.3). Except for technical limitations, all specimens tested with conventional methods yield a measurement and therefore both DSe and DSp can reach extreme values (i.e. 0% or 100%). For qPCR, DSp is expected to never reach 0% at the endpoint Ct value since most of D- individuals would not yield a

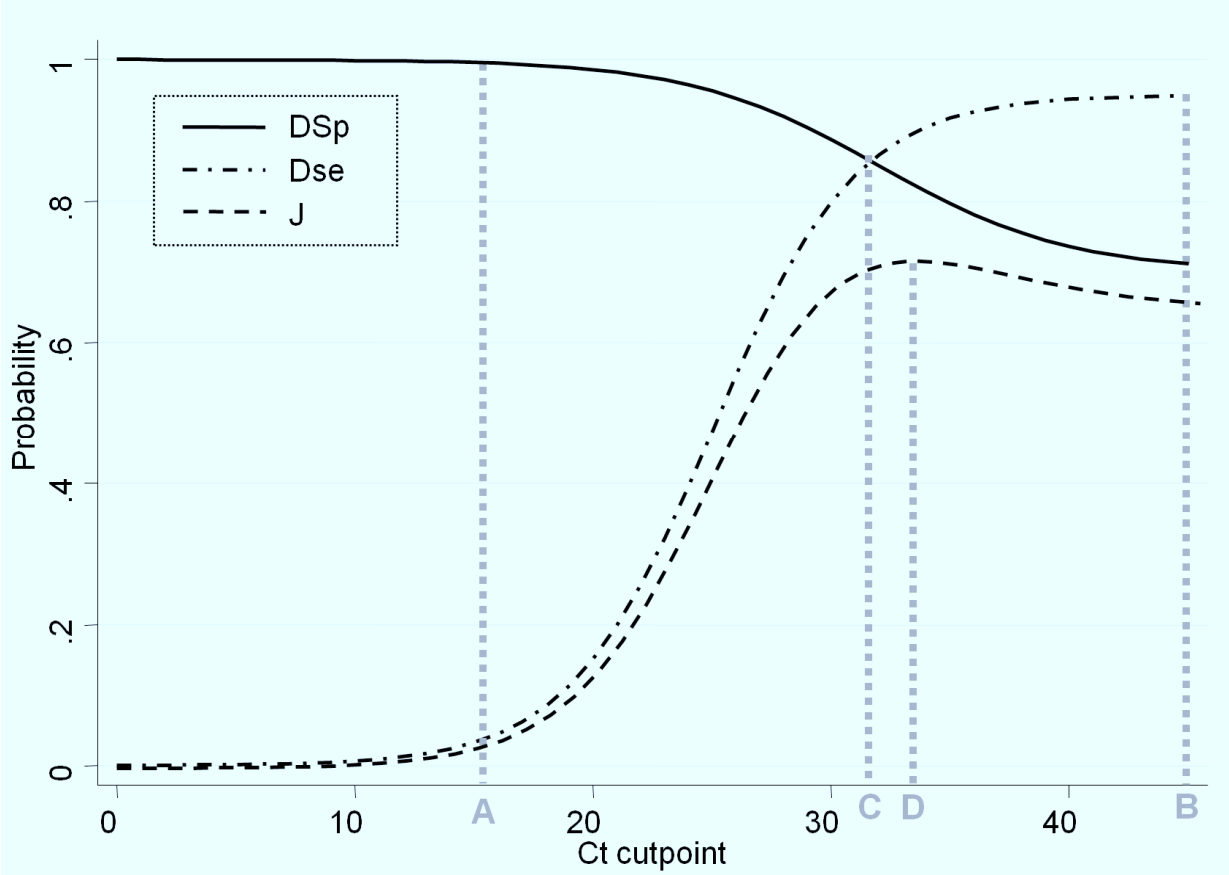


Fig. 5.3. Strategies to select a cycle threshold (Ct) cutpoint based on the probability of misclassification given the infection/disease status using a hypothesized Two-Graph Receiving Operating Characteristic (TG-ROC) curve.

Best DSp (line A); best DSe (line B); best combination of DSp/DSe (line C); best J (line D). DSp: diagnostic specificity; DSe: diagnostic sensitivity; J: Youden index (DSe + DSp -1).

Ct value. Similarly, DSe is not expected to reach 100% since false negatives, when they exist, do not yield a Ct value (Fig. 5.3).

When the true status of the sample is known (i.e. a gold standard available), automated non-parametric or parametric (assuming a specified distribution), computations to generate TG-ROC curves exist in varying statistical packages. For instance, the Stata statistical package (Stata Corp., College Station, TX, USA, 2007) offers a `roctg` command (Reichenheim, 2002). In the situation where the health status is unknown, Branscum et al. (2008) developed Bayesian estimations of semiparametric standard ROC curves that enable construction of the TG-ROC.

5.2.2.2 Minimization of the probability of misclassification

According to the World Organisation for Animal Health (OIE) requirements, to be validated for international trade, an assay must be evaluated to demonstrate its fitness for a specific purpose (OIE, 2009). The purpose of the test is to be defined first, then DSe and DSp are evaluated to assess if the test operating characteristics suit that intended use. The OIE lists six different purposes for using a test (OIE, 2009):

(i) Demonstration of freedom from infection in a defined population;

(ii) Confirmatory diagnosis of suspect or clinical cases;

(iii) Determination of immune status in individual animals or populations;

(iv) Certification of freedom from infection or agent in individual animals or products for trade purpose;

(v) Eradication of disease or elimination of infection from defined populations;and

(vi) Estimation of infection or exposure prevalence to facilitate risk analysis.

5.2.2.2.1 Probability of misclassification given the health status

A test will be validated and certified according to how well it performed for a specific purpose. For purposes (i), (ii) and (iii), the test validation will prioritize DSp (therefore PPV) over DSe. The Ct cutpoint is set such that DSp is maximized and DSe is optimized (Fig. 5.3, line A). For purpose (iv), the test validation will prioritize DSe (therefore NPV) over DSp. The Ct cutpoint is set such that DSe is maximized and DSp is optimized (Fig. 5.3, line B). Generally, for purpose (v), if the agent has zoonotic consequences for human safety, DSe (therefore NPV) will be prioritized over DSp, whereas, if the agent has mainly economic consequences, DSp (therefore PPV) will be prioritized over DSe. Finally, for purpose (vi), minimization of overall misclassification is the priority, regardless of the infection/disease status. This situation occurs when the test user intends to optimize the estimation of prevalence of infection/disease in a population. In this instance, the nature of misclassification (false positive or false negative) does not have an impact on the final purpose of the test as the estimated apparent prevalence will be corrected to obtain the true prevalence.

The most intuitive way to select a Ct cutpoint to minimize misclassification is to identify the Ct cutpoint that yields the best combination of DSe and DSp. That is, the Ct value for which DSe and DSp are equal or for which the square of the difference between the two parameters is minimized (Reichenheim, 2002). Practically, this cutpoint

corresponds to the Ct value where DSe and DSp curves intersect (Fig. 5.3, line C). This approach is used when no clear purpose of the test is predefined and operators desire a balanced performance for misclassification of both positives and negatives. This approach does not, however, minimize overall proportion of misclassification given the health status. The alternative approach is to select the cutpoint where J is maximized (Fig. 5.4, line A). Maximizing Youden's Index minimizes the sum of the false positive and false negative proportions. This approach may differ from the best pair approach (Fig. 5.3, line D).

5.2.2.2.2 Selection comparing ratio of correct and incorrect misclassification

LR^+ compares the proportion of correctly classified D+ individuals (true positive probability) over the proportion of misclassified D- individuals (false positive probability). The selection of a Ct cutpoint using LR^+ can follow two different approaches. First, a cutpoint can be set such that the probability to be incorrectly classified equals the probability to be correctly classified (i.e. $LR^+ = 1$). As a result, when the Ct value is below this cutpoint, the sample is deemed positive and the probability of a true positive is higher than the probability of false positive. According to Eq. (4), this is translated as DSe = DSp = 0.5 which is rarely a desirable case in practice. Therefore this approach is of little value and not recommended. Alternatively, a cutpoint can be set such that the probability to be correctly classified, compared to the proportion of incorrectly classified, is as large as possible. In this instance, the cutpoint is selected for the maximum corresponding LR^+ (Fig. 5.4, line B).

A similar approach is used for LR^-, whereby LR^- compares the proportion of

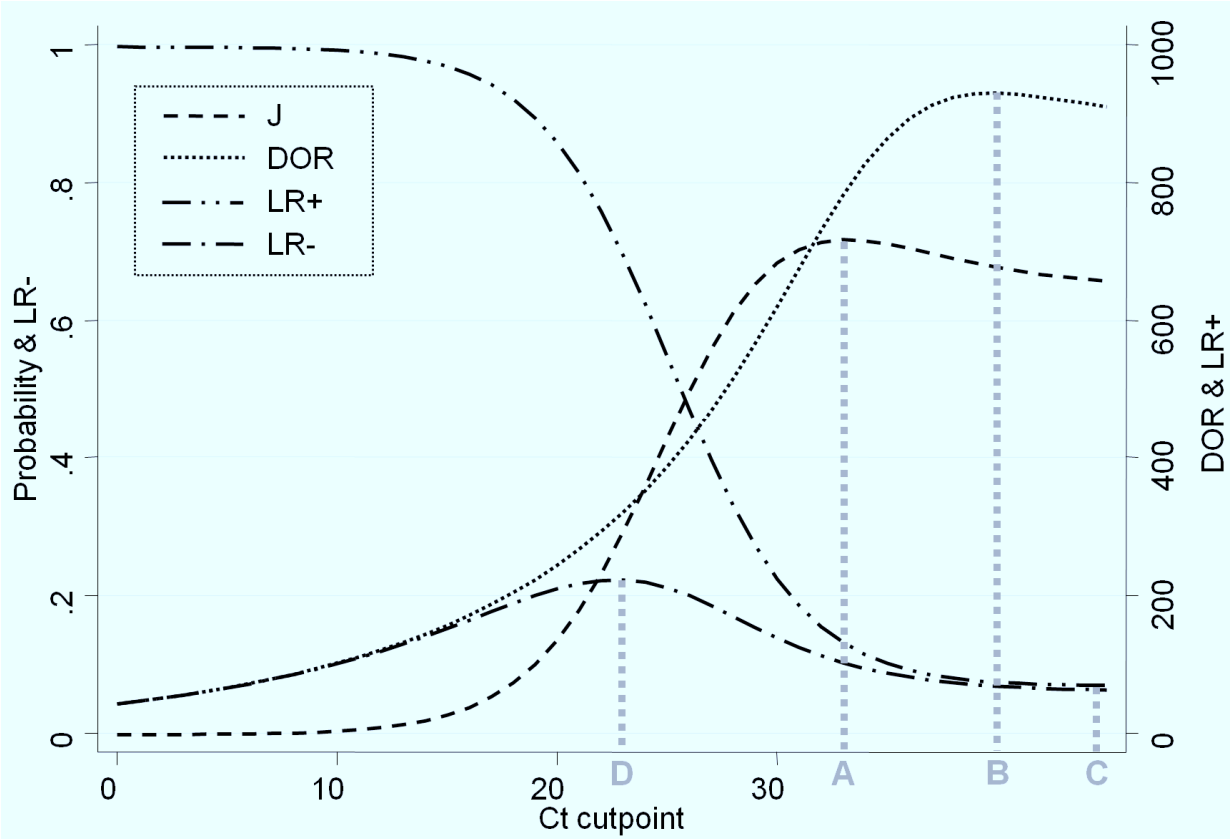
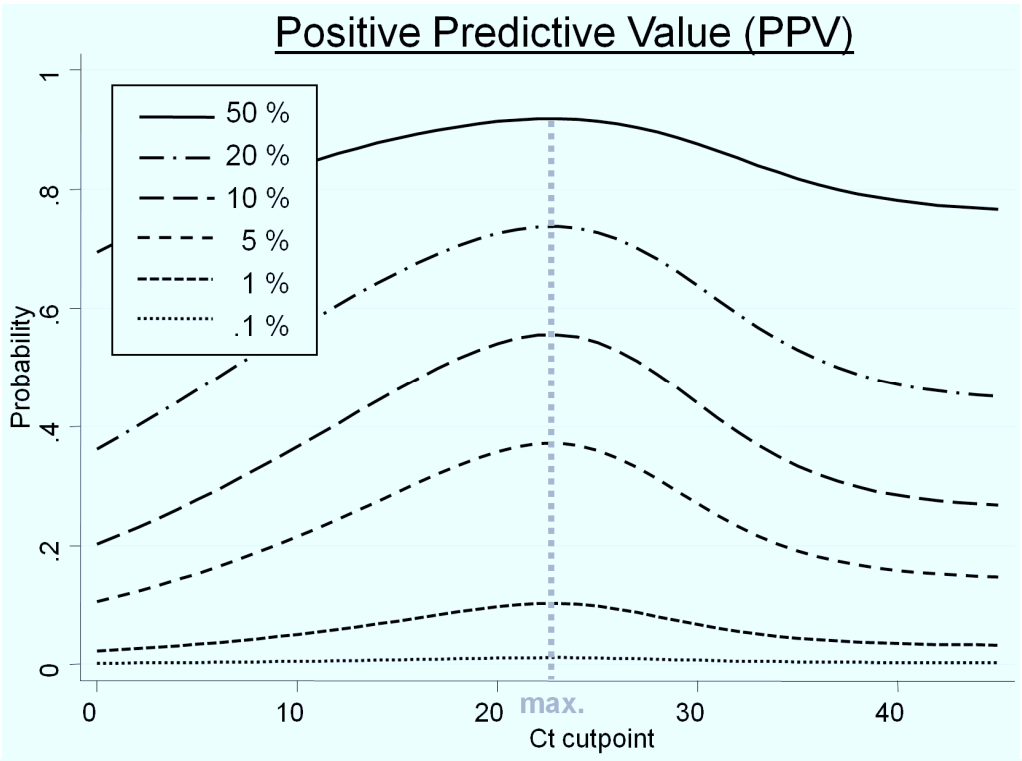


Fig. 5.4. Strategies to select a cycle threshold (Ct) cutpoint based on different probabilities of misclassification.
Best J (line A); best DOR (line B); best LR^+ (line C), best LR^- (line D). J: Youden index (DSe + DSp -1). DOR: diagnostic odds ratio; LR^+: likelihood ratio of a positive test; LR^-: likelihood ratio of a negative test.

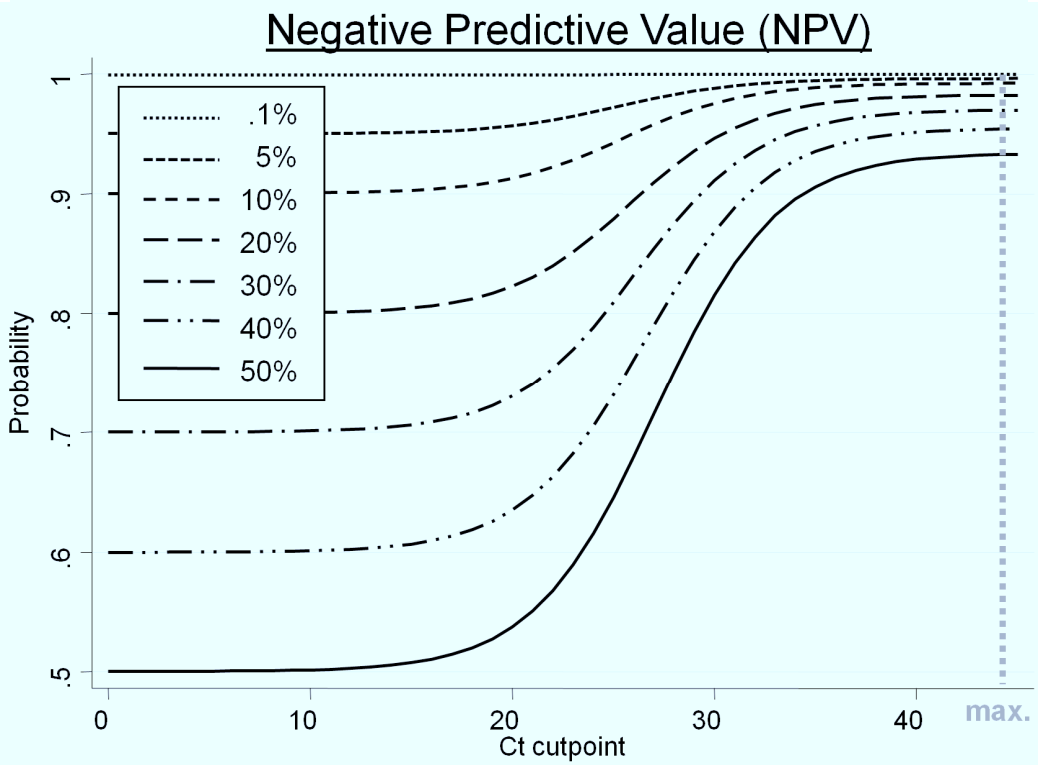
misclassified D+ individuals (false negative probability) over the proportion correctly classified of D- individuals (true negative probability). Then, the preferred cutpoint is selected where LR^- is minimal (Fig. 5.4, line C). The DOR combined both LR^+ and LR^-. In this instance, the Ct cutpoint is selected to maximize the chance of a correct classification relative to the chance of misclassification when DOR is maximal (Fig. 5.4, line D). DOR expresses how much more likely a sample is to be correctly classified than misclassified. Due to its symmetry, DOR can be interpreted either for a given test result or a given health status.

5.2.2.3 Probability of misclassification given a test result

The following probabilistic approach selects a Ct cutpoint to optimize the predictive value that depends on the assumed prevalence of D+ in the population. Similar to LRs, two approaches can be used. First, the cutpoint is set such that, above this limit, the probability of being truly classified (PPV) is lower than the probability of a false positive (complement of PPV, 1-PPV). This differs from the LR^+ approach since the given information is the test result and not the health status. This approach may appear as the most relevant since the true status is, in reality, unknown. The cutpoint corresponds to the Ct value where PPV = 1-PPV, which is equivalent to PPV = 50%. Since, depending on the prevalence, the predictive value might never cross the 50% limit (Fig. 5.5), cutpoints set where either both PPV and NPV are maximal (Fig. 5.5A and B, respectively) is a preferred approach. Based on the tested population, realistic prevalences should be hypothesized to optimize the selection of the cutpoint.



A.



B.

Fig. 5.5. Evolution of positive (A) **and negative** (B) **predictive values for corresponding infection prevalences and across cycle threshold** (Ct) **cutpoints** (max)**.**

DOR also expresses the ratio of the odds of PPV over the odds of NPV (Eq. (6)). Therefore, DOR combines both predictive value criteria and is conveniently independent of the prevalence. DOR is therefore a practical parameter that combines the optimization of all test operating characteristics at once regardless of the intended purpose.

5.2.2.4 Cost of misclassification

The cost of a misclassification is a relative concept that can refer to either the economic value (e.g. treatment cost) or the clinical and psychological consequences of a misclassification (e.g. fatal disease, treatment safety). Adapted from Vizard et al. (1990), a parameter "r" is used to relate the cost of misclassification in D+ (false negative cost) compared to D- (false positive cost) (r = false negative cost/ false positive cost). Depending on the context, misclassification costs are usually different ($r \neq 1$), although they can be equal (r =1).

5.2.2.4.1 Differing misclassification costs

In the misclassification cost approach, the Ct cutpoint will be selected such that the overall cost is minimized. The proportional cost of misclassification can be computed and monitored across Ct values using a formula derived from Anderson (1958):

$$Proportional\ cost = r * Pr * (1-DSe) + (1-P) * (1-DSp) \quad (7)$$

For instance, proportional costs were generated for a fixed prevalence (P = 10%) and using different values of r (Fig. 5.6). When r = 2 (cost of false negative is 2 time higher than the cost of a false positive), the curve clearly showed a minimum (Fig. 5.6, line A). When r < 1(cost of a false negative is inferior to the cost of a false positive), the curve only showed a minimum at Ct value = 0 when r = 0.5 (Fig. 5.6). For any given prevalence, the minimum of the proportional cost curve (cutpoint) moves towards the right (more sensitive test) as r increases (i.e. misdiagnosed D+ are becoming more costly).

5.2.2.4.2 Particular case of equal cost: Efficiency

According to Eq. (7), for r =1, the proportional cost equals P*(1-DSe) + (1-P)*(1-DSp). In this particular situation, the cost is the same for a false positive and a false negative. The monitored parameter now reflects the inefficiency (Inef) of the assay in contrast to the efficiency (Ef). The cutpoint is then set such that the Inef is minimized (or Ef maximized) (Fig. 5.6, line B).

In summary, the intended purpose of the test is critical in the selection process since Ct cutpoint may substantively change the operating characteristic of the test. Depending on the justification and the approach, several parameters may have to be considered (e.g. prevalence). However, DOR represents a good compromise to optimise the average test performance for positive and negative tests.

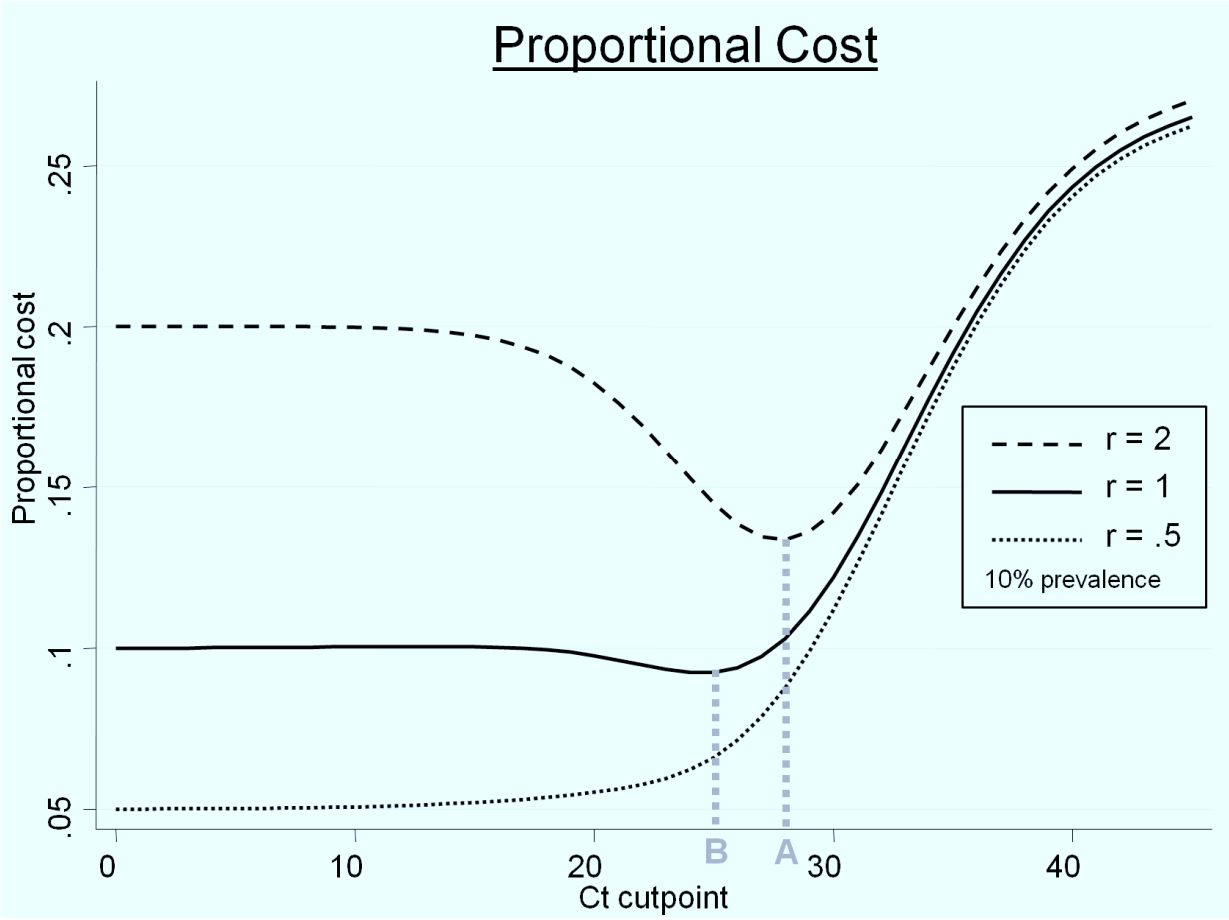


Fig. 5.6. Proportional cost across cycle threshold (Ct) **cutpoint for 10% prevalence of infected/diseased.** The parameter "r" represents the ratio of cost of a false negative over cost of a false positive. The Ct cutpoint is selected for the minimum proportional cost (lines A & B) in 2 situations for r.

5.3 Illustrations with an application

5.3.1 Background

A real-time reverse-transcriptase polymerase chain reaction (qRT-PCR) was recently developed and evaluated to detect infectious salmon anaemia virus (ISAV) in cultured Atlantic salmon in Canada (Chapter IV). The assay uses a Taqman probe in addition to a pair of primers to amplify a 120 base fragment of the 8^{th} RNA segment of the viral genome. Analytically, the fluorescence threshold was set by the computer software associated to the thermocyclor (MxPro QPCR Software, Stratagen). The reaction was run for a maximum of 45 cycles so that even a low copy number of target would be amplified and detected.

The initial purpose of this new test was to demonstrate freedom from infection (not disease) in a defined population. A total of 400 salmon were sampled from 4 different populations, and 112 samples generated Ct values ranging from 17.29 to 41.86 (Chapter IV). Interestingly, Ct values revealed a bimodal distribution suggesting two sub-populations of positive fish (Fig. 5.7). An estimation procedure without gold standard information (i.e. Latent Class Model) was run involving 4 other ISAV detection assays in a Bayesian framework (Chapter IV). The model identified three different classes of fish which were hypothesized as non-, low- and high-infected salmon (NI, LI and HI, respectively) (Chapter IV). The qRT-PCR was associated with two estimates of DSe (DSe_{LI} & DSe_{HI}), while DSp was unique (NI). The model was run for each possible qRT-PCR cutpoint along the range of obtained Ct values. For each CT cutpoint, three

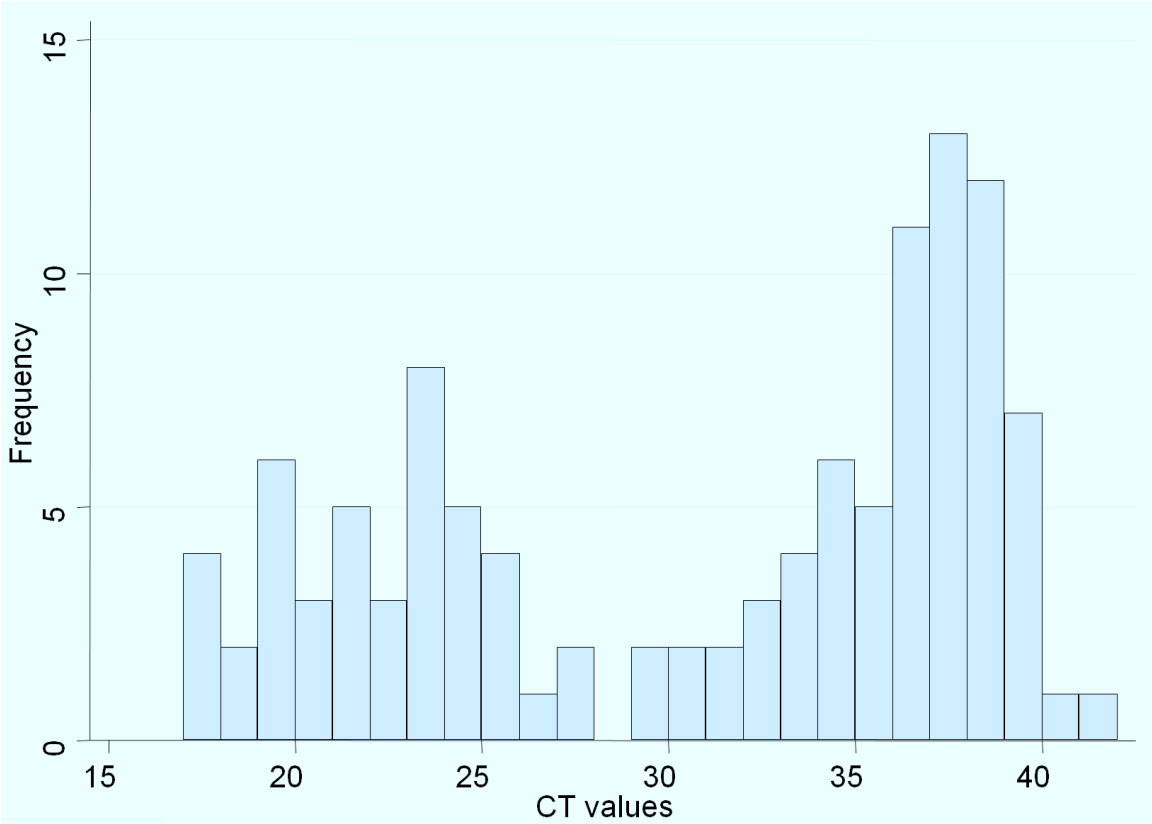


Fig. 5.7. Histogram of the cycle threshold (Ct) **values generated from 112 Atlantic salmon testing positive for infectious salmon anaemia virus** (ISAV) **with real-time RT-PCR.**

estimates (DSe_{LI}, DSe_{HI}, DSp) were obtained and plotted into a TG-ROC (Fig. 5.8). To facilitate the interpretation of this example, a combined DSe (DSe_C) was generated corresponding to weighted-averages of both DSe estimates as a function of the estimated proportions of LI and HI at each cutpoint (Fig. 5.8) and was used for all subsequent examples. In addition, we smoothed the curves to reduce noise due to statistical instability during the estimation process (Fig. 5.8).

5.3.2 Probabilistic approach

The original test purpose was the demonstration of freedom from infection in a salmon population. For trade purposes, an animal producer may be required to prove that the products are free of a specified pathogen. The initial concern was to minimize false positives because of substantial economic consequences when a population is falsely declared not free.

Three different approaches can be used by the test operator to select a Ct cutpoint. To ease the process, a TG-ROC was generated for J, DOR, LR^+ and LR^- computed from either the exact (Fig. 6.9A) or the smoothed estimates of DSe_C and DSp (Fig. 6.9B). Smoothing intended to remove the statistical noise, however, it modified the curves profile and the subsequent interpretation of the cutpoints when curves are not monotonic. Corresponding with the reality of the data, we selected the cutpoints based on exact estimates profiles. First, the test operator wants to minimize the probability of false positives among infected salmon, which corresponds to the maximum DSp associated

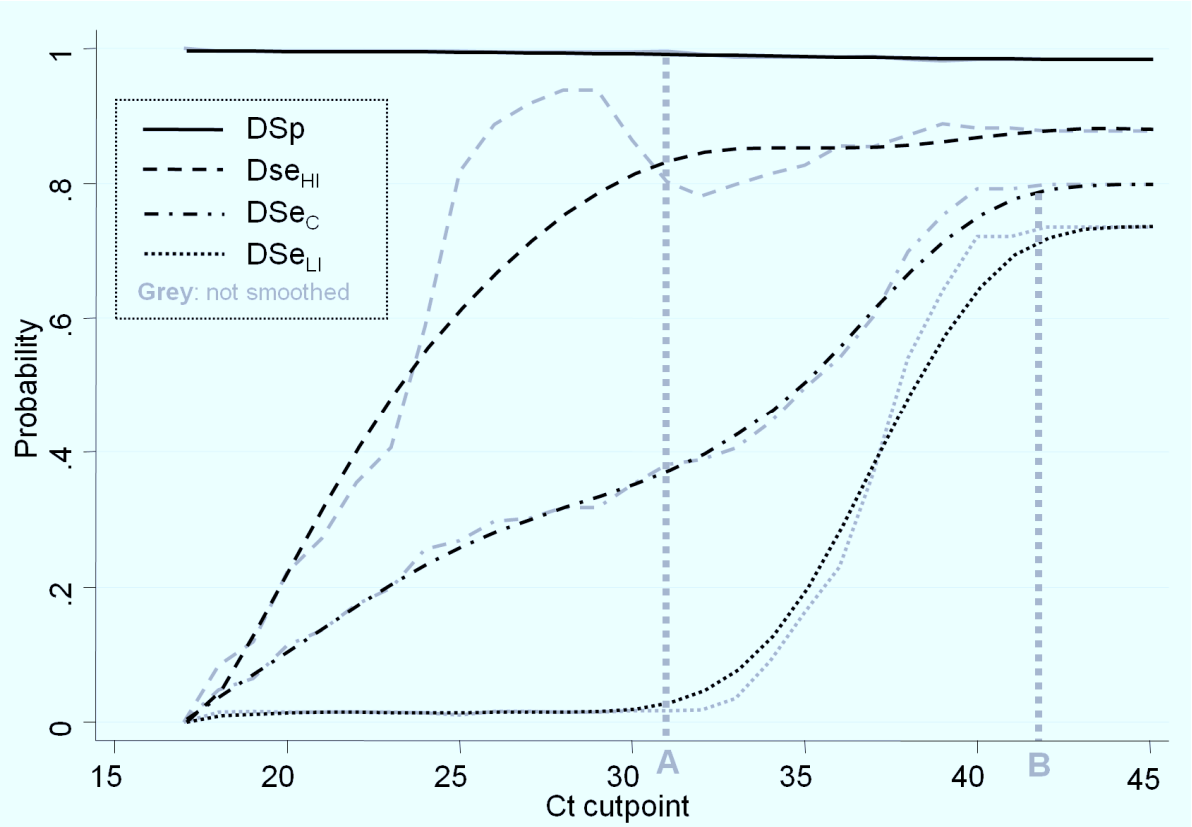
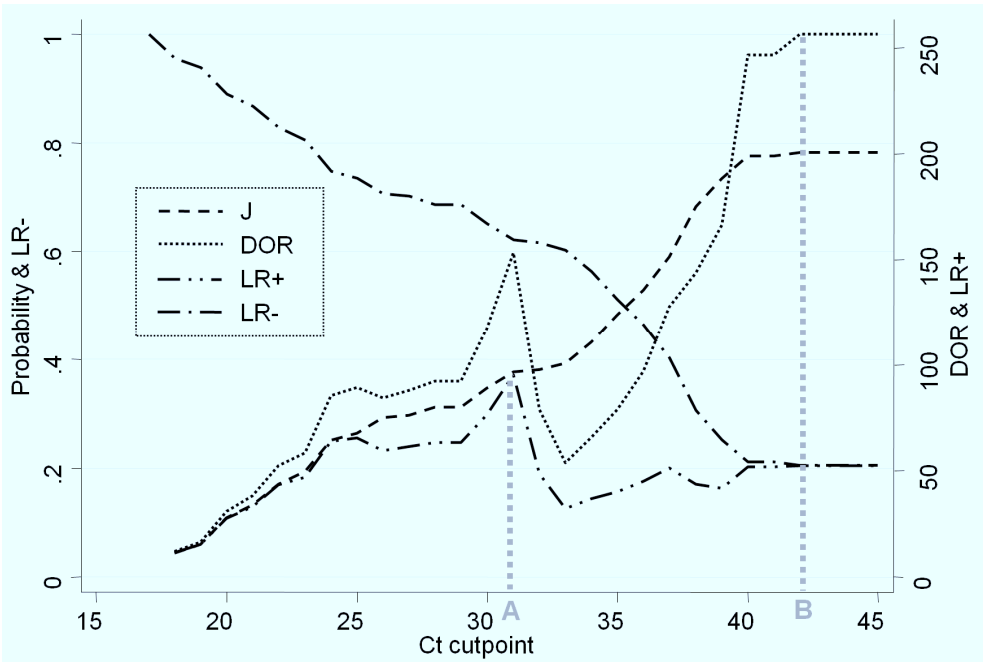


Fig. 5.8. Two-Graph Receiving Operating Characteristic (TG-ROC) **curve estimated for a real-time RT-PCR assay for infectious salmon anaemia virus** (ISAV) **with selection of cutpoints for best DSp** (line A) **and best combined DSe$_C$** (line B)**.** Grey curves show original estimations that were smoothed in black. DSp: diagnostic specificity; DSe$_{HI}$: diagnostic sensitivity in highly infected fish (HI); DSe$_{LI}$: diagnostic sensitivity in lowly infected fish (LI); DSe$_C$: combined diagnostic sensitivity.

A.



B.

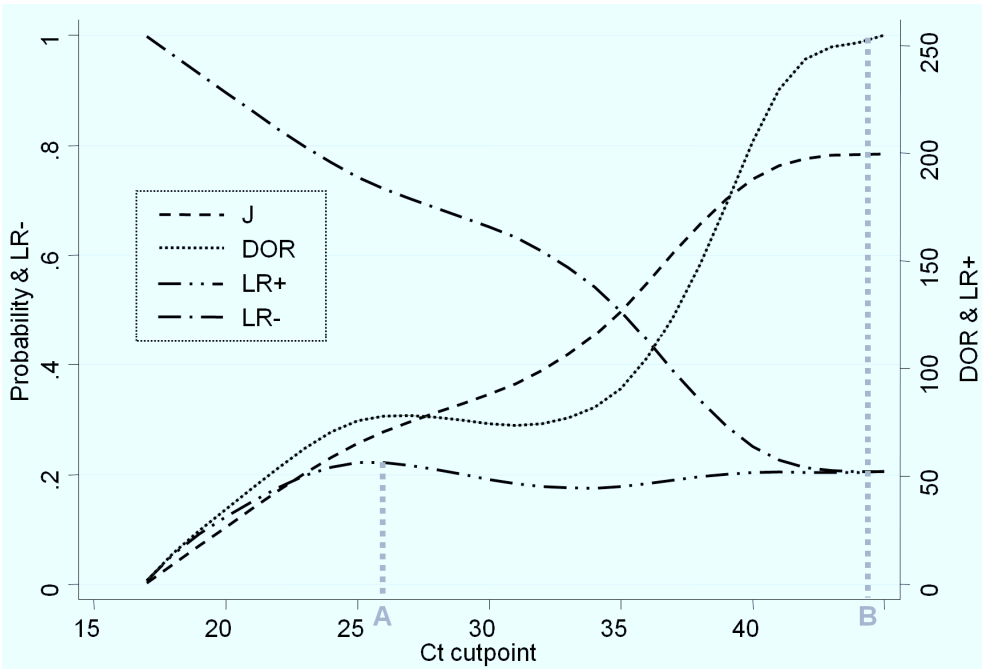
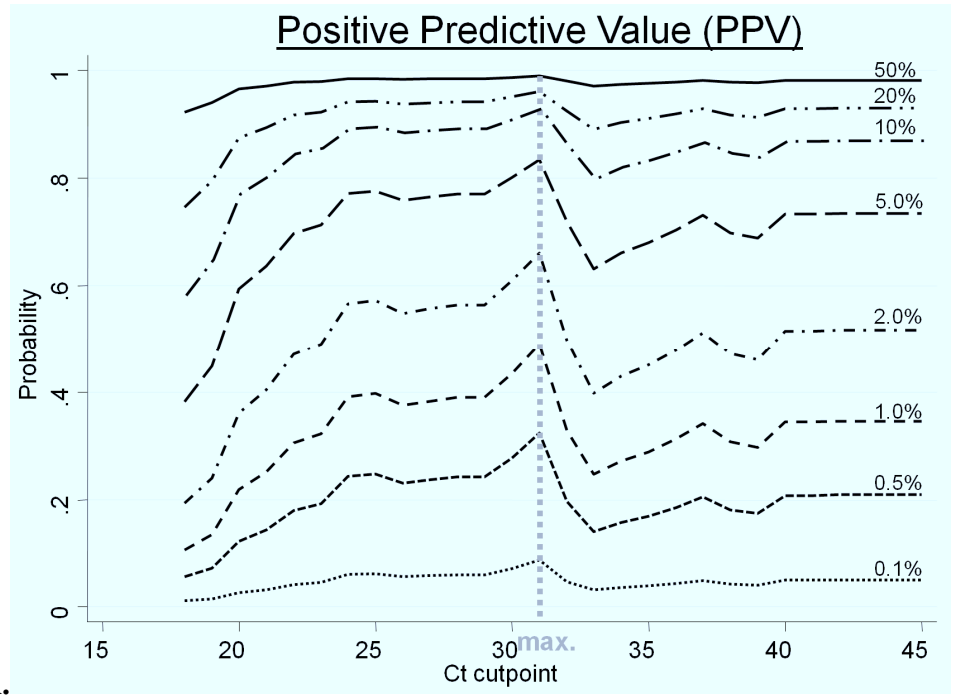


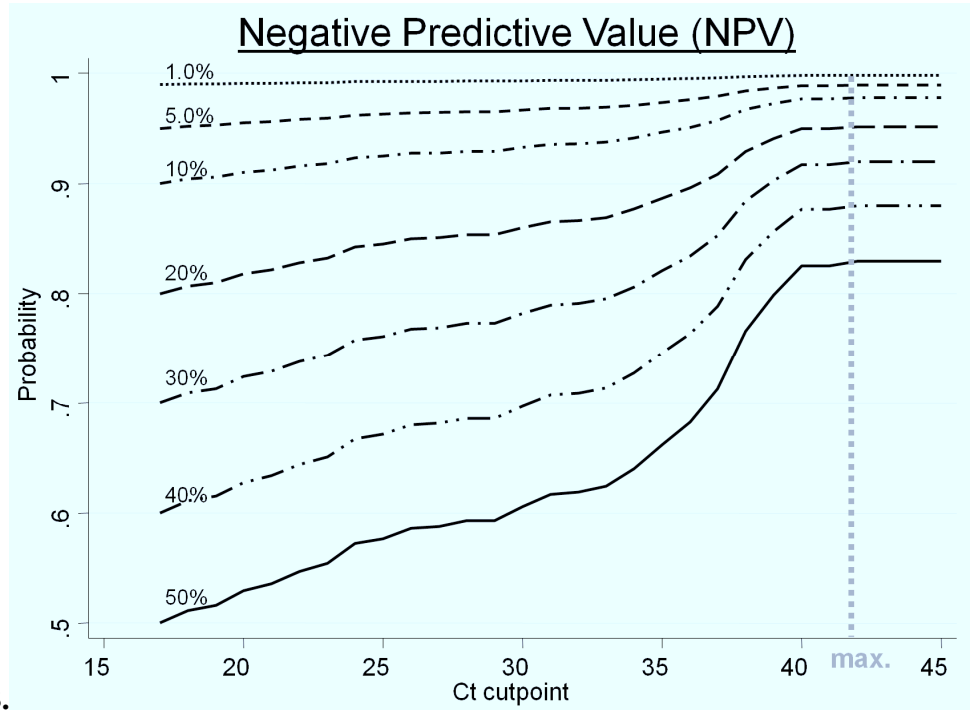
Fig. 5.9. Selection strategies to select a cycle threshold (Ct) **cutpoint based on different single misclassification parameters of test accuracy.** Estimated curves are based on exact estimates of DSe and DSp (A), and based on smoothed estimates of DSe and DSp (B). Best LR^+ (line A); best J, DOR and LR^- (line B). DSp: diagnostic specificity; DSe: diagnostic sensitivity; J: Youden index (DSe + DSp -1); DOR: diagnostic odds ratio; LR^+: likelihood ratio of a positive test; LR^-: likelihood ratio of a negative test.

with an acceptable DSe (subjective assessment) (Ct = 31, Fig. 6.8, line A). Alternatively, the test operator desires to maximize the probability of true positives among infected fish relative to the probability of false positives among non-infected fish (i.e. maximal LR^+) (Fig. 6.9, line A). Finally, the operator wants to minimize the probability of false positives among positive tests. In this instance, the cutpoint is selected for the maximum PPV and, regardless of the prevalence, the maximum PPV is reached at the same cutpoint as previously (Ct = 31, Fig. 6.10A). Interestingly, the different approaches selected the same cutpoint. The objective of minimizing false positives may also be compatible when seeking the confirmation of a clinical or suspect case, the determination of the immune status of an individual, or the eradication of infection associated with economic impact (see section 5.2.2.2).

Nonetheless, the "*absence of evidence is not evidence of absence*" and disregarding all results between 31 and 45 cycles may minimize false positives but ignores many probable true positives. Therefore, a buyer may want to screen the products to verify the absence of infection. In this instance, the operator would prioritize the minimization of false negatives to avoid the introduction of the pathogen. Similarly, these alternative approaches include maximization of DSe to minimize the probability of false negatives among infected fish (Fig. 5.8, line A), or minimization of LR^- to minimize the probability of false negatives among infected fish relative to the probability of true negatives among non-infected fish (Fig. 5.9, line B), or maximization NPV to minimize the probability of false negatives among negative tests (Fig. 5.10B). Regardless of the approach, identical cutpoints are selected (Ct= 42). No Ct values were observed above 42 cycles



A.



B.

Fig. 5.10. Evolution of positive (A) **and negative** (B) **predictive values for corresponding infection prevalences and across cycle threshold** (Ct) **cutpoints of a real-time RT-PCR assay for infectious salmon anaemia virus** (ISAV).

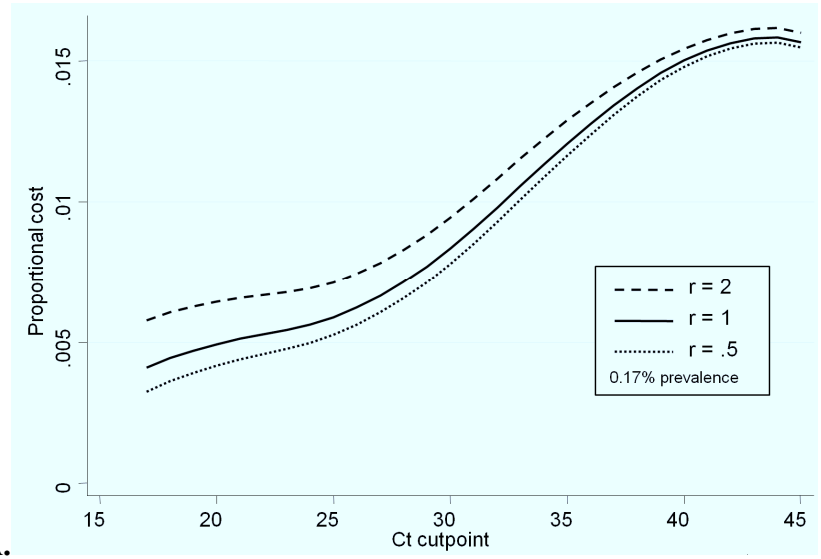
which provided an end plateau. Believed to be representative of the full range of ISAV infection in true salmon population, it seems unlikely that infected salmon would yield Ct above this limit and the reaction could therefore be ended earlier at 42 cycles. Amplification efficiency may, however, vary across runs and a same sample could yield different Ct values (estimated repeatability = ±2 cycles), making 45 cycles a reasonable endpoint.

In addition, to satisfy both trade partners, the test operator may want to optimize the overall performance of the assay. The most intuitive approach is to select the best combination of DSe and DSp occurring when both curves cross each other. Unfortunately, the two curves do not intersect and the best combination is obtained when DSe is maximal at Ct = 42 (Fig. 5.8A). Otherwise, the operator may select a cutpoint to minimize overall misclassification regardless of the animal health status. First, by minimizing the proportion of misclassification, a cutpoint is selected for maximum J (Ct = 42, Fig. 5.8B). Alternatively, a cutpoint might be selected for the maximum DOR (proportion of correctly classified compared to misclassified samples). The corresponding cutpoint is again 42 cycles.

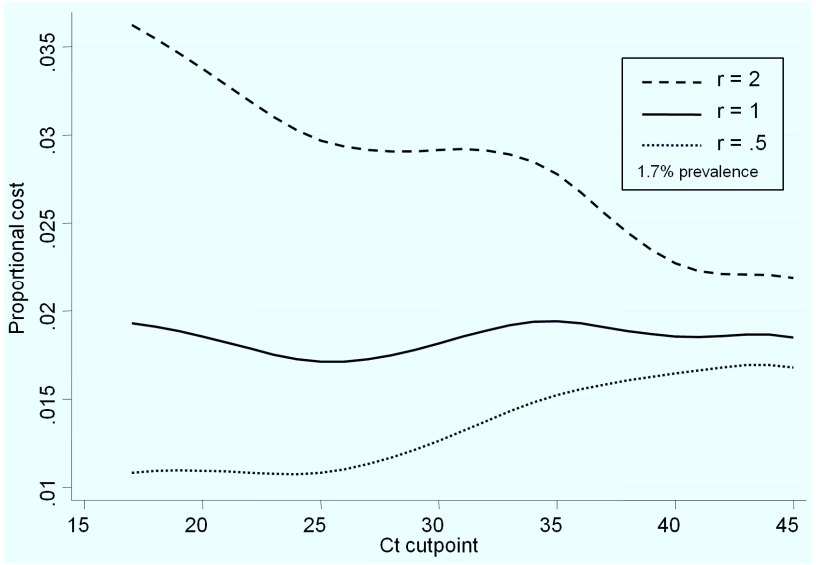
The majority of the above mentioned justifications (minimal false negatives or overall misclassification) strongly support 42 cycles as the recommended cutpoint. The initial objective was, however, to limit or avoid the economic burden associated with false positive results and as a result 31 cycles would be a more appropriate choice.

5.3.3 Cost approach

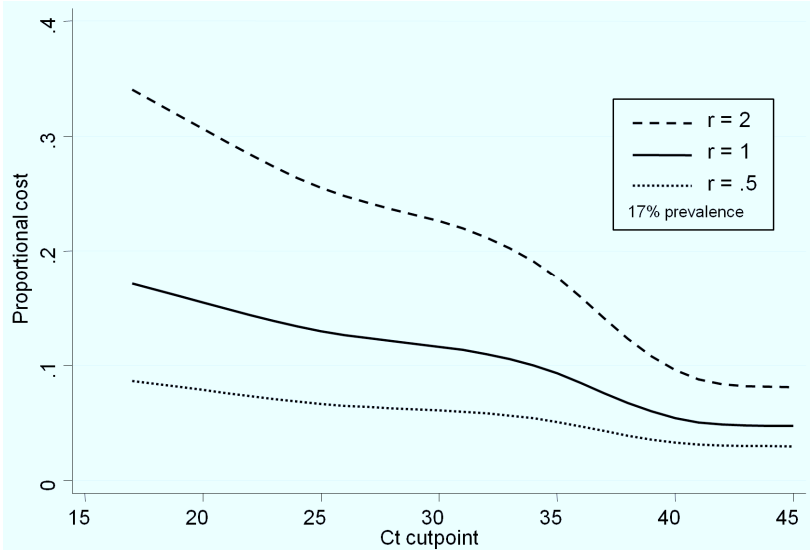
The initial objective of demonstrating freedom from infection in a population implies that the cost of a false positive is more important than the cost of a false negative. As a consequence, for this approach, the relative cost of misclassification in D+ (false negative) compared to D- (false positive) was set at half (r = 0.5). In addition, the reverse scenario (r = 2) and the equal cost (r = 1) options were also investigated. Associated to r, DSe and DSp, the proportional cost of misclassification also depends on prevalence (Eq. (7)). As the prevalence was expected to be fairly low when demonstrating freedom from infection, 3 low prevalences were included in the simulation: 0.17%, 1.7%, 17% (Fig. 5.11A, B & C, respectively). All curves were monotonic and none demonstrated an intermediate minimum. At 0.17% prevalence, regardless of the relative cost (r), the lowest proportional cost was obtained for the minimum observed Ct value (i.e. 17 cycles). This was explained by the fact that at such a low prevalence, most of the positive results obtained were highly likely to be false. At 1.7% prevalence, the Inef of the test (r = 1) was stable due to a constant probability of a false test result regardless of the Ct value. This may be explained by a progressive increase of false positives, and a reciprocal decrease of false negatives, as the cutpoint increases. The proportional cost was therefore increased when the relative cost of false positives increased (r = 0.5), and proportional cost decreased when the relative cost of false negatives increased (r = 2) Fig. 5.11B). At 17% prevalence, the probability of false negatives increased when the cutpoint increased. Therefore, regardless of the relative cost, the proportional cost progressively decreased, establishing the largest obtained Ct value as the cutpoint (i.e. Ct =42). At a low



A.



B.



C.

Fig. 5.11. Proportional cost across cycle threshold (Ct) **cutpoint for three prevalence levels** (0.17% (A), 1.7% (B) and 17% (C)) **of infected fish**. The parameter "r" represents the ratio of cost of a false negative over cost of a false positive, and the Ct cutpoint is selected for the minimum r. Exact proportional cost curves were smoothed.

prevalence, none of the positive test results can be verified and confirmed in the field, regardless of the cutpoint. As a result, the selection of a cutpoint would not really change the test performance since the probability of detection is dominated by the very low probability to sample an infected fish.

5.4 Discussion

Although TG-ROC curve is a convenient visual tool to select a cutpoint for real-time PCR, interpreting the profiles should be done cautiously. DSe and DSp are strongly influenced by the population-level distribution of biological factors associated with the pathophysiology of the disease (e.g. age, gender, infection stage) (Greiner and Gardner, 2000). In this instance, the profile of DSe curves in the TG-ROC is expected to substantially change according to the stage of ISAV infection in a salmon. Since the distribution of infection stages (associated with the viral load and therefore Ct values) varies across populations, the TG-ROC profile is population-dependent and should be validated for specific targeted populations. Similarly, false positive results that yield a certain Ct value might be explained, for instance, by cross-contamination (Wilson, 1997). Therefore, the profile of the DSp curve might also change dramatically due to increased contamination pressure when handling tissues from heavily infected populations. The representativeness of specimens used to construct the TG-ROC and the strong dependence of curve profiles on the tested population should be of primary consideration when selecting a cutpoint (Greiner & Böhning, 1994). The methods used to estimate the TG-ROC curves may be complex and require the use of advanced statistical models (Branscum et al., 2008). The use of

experimentally challenged animals to evaluate test operating characteristics is not recommended since the specimens are not representative of the infection spectrum in field populations and parameters are often overestimated (OIE, 2009).

The selection of a cutpoint for real-time amplification assays relies on interpretation of test results (Ct value) that can be justified based on a variety of factors. The purpose of the test and the parameters of interest should, however, be clearly stated prior to initiating the evaluation to assist the identification of the most appropriate cutpoint. For instance, for demonstration of freedom from infection/disease, the primary objective would be to minimize the probability or the cost of a false positive. Corresponding DSe (or NPV) can, however, be quite low and limit the detection of infected specimens, suggesting that other parameters, such as DOR, may be of greater interest.

Theoretically, it would be rare to observe the DSe and DSp curves intersecting for real-time amplification assays. Indeed, as found in the example, DSp (especially with probe-based fluorophore chemistry) only decreased moderately, and DSe did not reach perfection (100%) preventing the intersection. An easy way to predict the presence or absence of intersection is to compare the estimates of DSe and DSp when no Ct cutpoint is set. When DSe is higher than DSp, an intersection is expected; alternatively, when DSe is lower than DSp, no intersection is expected.

Predictive values (PPV & NPV) depend on prevalence. However, their respective maxima rely only on the test DSe and DSp and are fixed regardless of the prevalence. Conversely, the proportional cost of misclassification is highly influenced by prevalence, such as when DSe is not perfect, the proportion of false negative results increases with prevalence.

5.5 Conclusion

Several analytical and epidemiological approaches exist to select a cutpoint for real-time amplification assays. Often, the same Ct cutpoints may be selected, despite the use of different approaches and justifications. The epidemiological cutpoints are population-dependent and their validity is directly associated with the targeted population. Based on our example study on ISAV in salmon, even if the use of a cutpoint is reasonable, it is still recommended to report the Ct value to the end decision-maker with complementary analytical or epidemiological information on the test performance to justify the classification of infection/disease status. Finally, real-time amplification efficiency varies within and between laboratories, and a set cutpoint might not be constant across runs. Therefore, when used, it is recommended to normalize cutpoints with relative or absolute quantification approaches.

5.6 References

Alberg, A.J., Park, J.W., Hager, B.W., Brock, M.V., Diener-West, M., 2004. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J. Gen. Intern. Med. 19, 460-465.

Ambruster , D.A., Tillman, M.D., Hubbs, L.M., 1994. Limit of detection (LOD)/ limit of quantification (LOQ): comparison of the empirical and statistical methods exemplified with GC-MS assays of abused drugs. Clin. Chem. 40, 1233-1238.

Anderson, T.W., 1958. An introduction to multivariate statistical analysis. Wiley, New York, pp. 126-131.

Branscum, A.J., Johnson, W.O., Timothy, E.H., Gardner, I., 2008. Bayesian semiparametric ROC curve estimation and disease diagnosis. Stat. Med. 27, 2474-2496.

Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat. Med. 16, 981-991.

Burns, M., Valdivia, H., 2008. Modelling the limit of detection in real-time quantitative PCR. Eur. Food Res. Technol. 226, 1513-1524.

Dohoo, I., Martin, W., Stryhn, H. (Eds.), 2009. Veterinary Epidemiologic Research. 2nd ed., AVC Inc., Charlottetown, Canada.

Glas, A.S., Lijmer, J.G., Prins, M.H., Bonsel, G.J., Bossuyt, P.M., 2003. The diagnostic odds ratio: a single indicator of test performance. J. Clin. Epidemiol. 56, 1129-1135.

Greiner, M., Böhning, D., 1994. Letter to the editor: Notes about determining the cut-off value in enzyme linked immunosirbent assay (ELISA)- Reply. Prev. Vet. Med. 20, 307-310.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 42, 2-22.

Greiner, M., Sohr, D., Göbel, P., 1995.A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. J. Immunol. Methods. 185, 123-132.

Mackay, I.M., Arden, K.E., Nitsche, A.. 2002. Real-time PCR in virology. Nucleic Acids Res. 30, 1292-1305.

Mehra, S., Hu, W.S., 2005. A kinetic model of quantitative real-time polymerase chain reaction. Biotechnol. Bioeng. 91, 848-860.

Office International des Epizooties, 2009. Manual of Diagnostic Tests for Aquatic Animals 2009. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 10-30.

Rebrikov, D.V., Trofimov, D.Yu., 2006. Real-time PCR: a review of approaches to data analysis. Prikl. Biokhim. Mikrobiol. 42, 520-528.

Reichenheim, M.E., 2002. Two-graph receiver operating characteristic. Stat J. 2, 351-357.

Rutledge, R.G., 2004. Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. Nucleic Acids Res. 32, e178.

Snow, M., McKay, P., Matejusova, I., 2009. Development of a widely applicable positive control strategy to support detection of infectious salmon anaemia virus (ISAV) using Taqman real-time PCR. J. Fish Dis. 32, 151-156.

Sunderman, F.W.Jr., 1975. Current concepts of "normal values," "reference values," and "discrimination values," in clinical chemistry. Clin. Chem. 21, 1873-1877.

Vecchio, T.J., 1966. Predictive value of a single diagnostic test in unselected populations. N. Engl. J. Med. 274, 1171-1173.

Vizard, A.L., Anderson, G.A., Gasser, R.B., 1990. Determination of the optimum cut-off value of a diagnostic test. Prev. Vet. Med. 10, 137-143.

Wilson, I.G., 1997. Inhibition and facilitation of nucleic acid amplification. Appl. Environ. Microbiol. 63, 3741-3751.

Wong, M.L., Medrano, J.F., 2005. Real-time PCR for mRNA quantitation. Biotechniques 39, 75-85.

Yerushalmy, J., 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Publ. Health Rep. 62, 1432-1449.

Youden, M.H., 1950. Index for rating diagnostic tests. Cancer 3, 32-35.

Chapter VI: GENERAL CONCLUSION

6.1 Introduction

The central theme of this thesis was to assess the influence of factors associated with the pathophysiology of ISAV infection on diagnostic test accuracy. Change in test operating characteristics according to population factors, such as health status (i.e. prevalence) and disease spectrum (i.e. degree of severity) make knowledge of these impacts necessary for understanding the performance of the test. Additional biological covariates may be associated with the manifestation of the infection/disease in the population, and influence test performance. Studies using field-based samples are necessary to properly represent the mixture distribution of these covariates and provide assessments applicable to the range of disease stages naturally occurring in target populations. Specifically, the use of field specimens is preferred as they reflect the true submission conditions (including the full sampling process) as opposed to bench studies where, for instance, molecular matrices are often artificial, diluted, or introduced.

The Standards for Reporting of Diagnostic Accuracy (STARD) recommended that variation of diagnostic performance be reported using different true populations (Bossuyt et al., 2003). Cross-sectional sampling ensures a good representation from a defined target population and allows unbiased extrapolations from the study population (i.e. internal validity). Except in a few conditions (e.g. chronic diseases), diseases progress and change in a population over time. For instance, in aquatic food animal production, each component of the "host-pathogen-environment" interaction varies

substantially, resulting in different manifestations of the disease. In this situation, results of a diagnostic test evaluation using a target population at one point in time might not be comparable to future or previous test results for that population at another point in time (i.e. internal validity), or other populations at other points in time (i.e. external validity). Although continuous monitoring of diagnostic test accuracy is a possibility, the repeated evaluation of detection methods is unrealistically labour-intensive and expensive. Alternatively, covariate-specific estimates can be obtained using stratification or modelling analyses, and subsequently used to predict test accuracy for projected mixture distributions of future or external populations.

This thesis focused on alternative methodologies for diagnostic test evaluation, including alternative methods for description and reporting of results, covariate-specific estimations, and prediction of operating characteristics.

6.2 Visualisation and reporting of descriptive diagnostic test studies

6.2.1 Screening of paired test results (Chapter II)

Before considering the implementation of covariate-specific estimation, screening of the test results obtained from the same specimens (paired design) are conducted to identify the presence of variant patterns associated with population-level factors. The traditional organisation of paired test results (samples in rows and tests in columns) was inverted and the results presented in a DNA-like alignment format (samples in columns

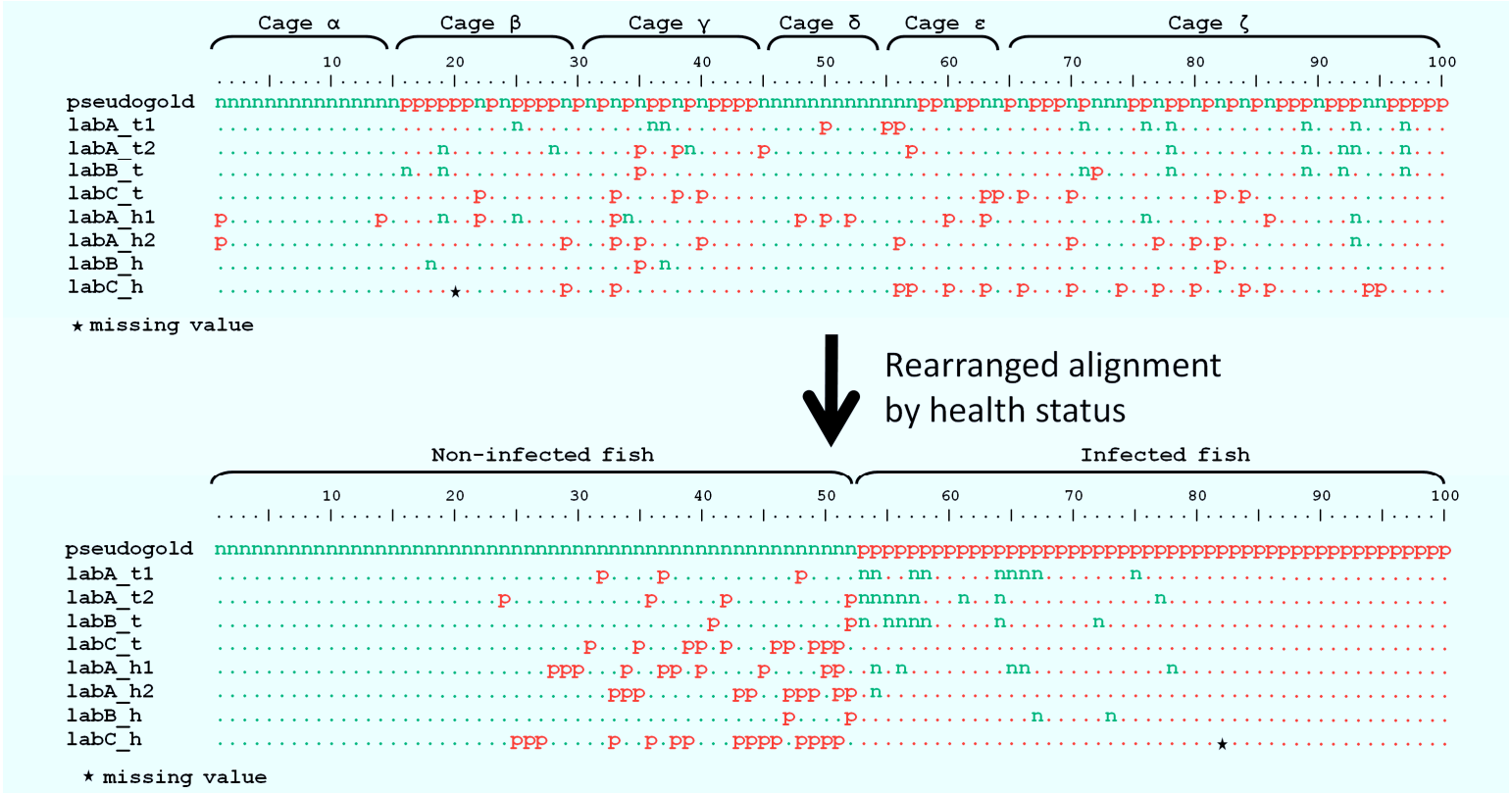


Fig. 6.1. Rearrangement of test result alignment according to the health status.
The top alignment corresponds to the comparison of results from 8 different test runs on 100 salmon, clustered by cage of origin (800 test results in total). The "pseudogold" corresponds to the true assumed infection status of the fish. In the bottom alignment, the salmon were rearranged and clustered by infection status. Two separate patterns of agreement between infected and non-infected fish (e.g. more red "p" than green "n") are revealed and a secondary factor should be considered to further explain them.

and tests in rows). The resulting test alignment provides a convenient and intuitive way to view and quickly compare test results by only highlighting discrepancies. By clustering the columns (specimens) by covariates, a particular testing pattern becomes evident to the examiner (Fig. 6.1). This process provides a compact, summarized and complete dataset format which is readily and easily transferrable to publication.

6.2.2 Full description of agreement (Chapter II)

Traditional studies of agreement include comparisons among detection methods to assess comparability, or comparisons among runs using the same method to assess repeatability and reproducibility. Each test is compared to another test and the resulting pair-wise measures of agreement are summarized in matrix table format. Although detailed information about a particular test comparison is available, the detection of a cluster of tests generating similar results becomes problematic, especially when numerous tests are involved.

Based on phylogenetic studies, we adapted a cluster-type analysis resulting in a tree-shaped diagram which grouped the tests for ISAV in salmon with substantial agreement (Fig 6.2). Using this method, the agreement between two tests can also be visually assessed by the relative length of the connecting branches (shorter branches coincided with greater agreement). When the number of discrepancies among tests increased, the phylogram had more discriminatory information and therefore more resolution to separate branches of tests. A set of tests that strongly agree will result in a low resolution tree in which branch nodes are close together (i.e. low bootstrap values

Distance	LabA_t1	LabA_t2	LabA_h1	LabA_h2	LabB_h	LabC_h	LabB_t	LabC_t
LabA_t1	\	**0.16**	**0.19**	**0.19**	**0.14**	**0.25**	**0.13**	**0.22**
LabA_t2	0.84	\	**0.23**	**0.19**	**0.14**	**0.25**	**0.09**	**0.2**
LabA_h1	0.81	0.77	\	**0.2**	**0.19**	**0.22**	**0.22**	**0.19**
LabA_h2	0.81	0.81	0.8	\	**0.11**	**0.14**	**0.18**	**0.13**
LabB_h	0.86	0.86	0.81	0.89	\	**0.19**	**0.11**	**0.12**
LabC_h	0.75	0.75	0.78	0.86	0.81	\	**0.24**	**0.15**
LabB_t	0.87	0.91	0.78	0.82	0.89	0.76	\	**0.19**
LabC_t	0.78	0.8	0.81	0.87	0.88	0.85	0.81	\

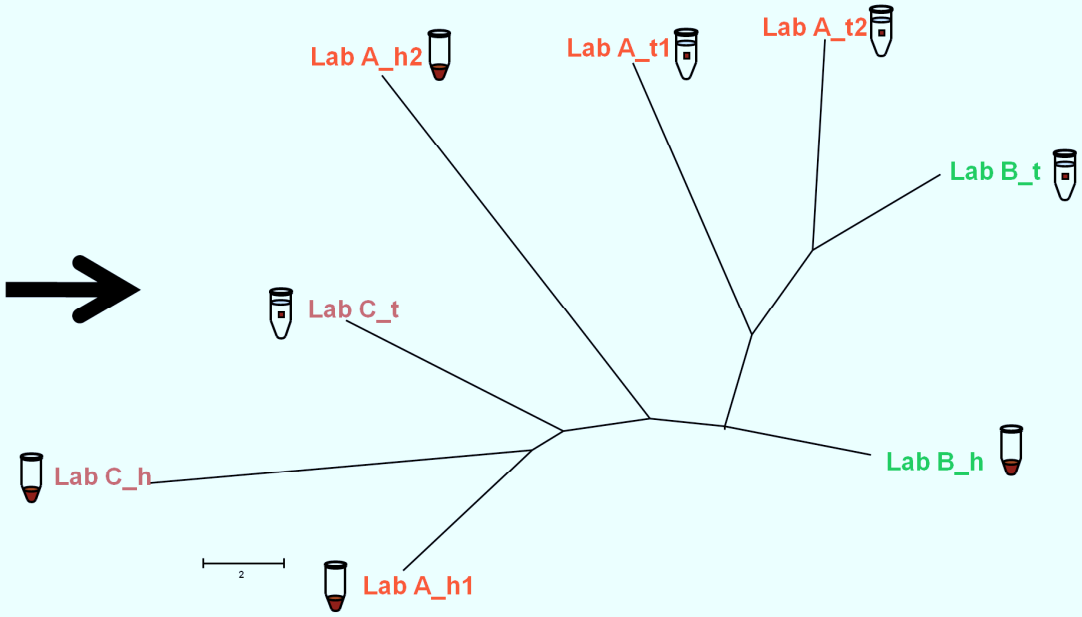


Fig. 6.2. Parallel comparisons between a test agreement matrix and a test phylogram.
The agreement matrix with proportions of agreement (lower left corner) and proportion of disagreement or distance (top right corner in bold) between runs is visually converted into a star-shaped, unrooted phylogram representing relative agreement among tests. The distance between two runs is visually assessed by the relative length of branches that connect them and are scaled based on the number of differing results.

giving weak support for separate branches) and no cluster information is identifiable. In the tree construction process, it was assumed that the weight associated with the probability of an infected/diseased (D+) animal testing negative (i.e. false negative fraction) was equal to the weight associated with the probability of non-infected/non-diseased (D-) animal testing positive (i.e. false positive fraction). This assumption may not be valid since diagnostic sensitivity (DSe) and diagnostic specificity (DSp) are rarely the same. Future improvements to this approach could be accomplished by assigning different weights to test result changes (i.e. from positive to negative and from negative to positive).

6.3 Covariate-specific estimation

6.3.1 Repeatability and reproducibility (Chapter III)

Although Yerushalmy (1947) initially expressed test trueness as two separate parameters, DSe and DSp, agreement continues to be expressed as a single overall parameter regardless of the animal health status. Expecting different degrees of agreement in D+ and D-, the concept of separate estimates for proportion of agreement (Pa) for each health status was introduced for ISAV tests in salmon. Estimation of Kappa in only D+ or only D- animals is, however, of little interest since prevalences in these two groups are extreme (100% and 0% respectively) and therefore erode the stability of Kappa. Similar to DSe and DSp, direct estimation procedures of specific agreement

require the true health status to be known (i.e. gold standard or pseudogold standard). The specific estimates of agreement are therefore estimated by simple stratification or modelling. Future research should be considered surrounding the development of modified latent class models (LCM) to estimate specific agreement without reference information on the health status.

Other factors, such as degree of infection, may be of interest when estimating specific agreement. During the assay development, the impact of other factors should initially be investigated analytically by assessing robustness. For instance, the evaluation of the effect of analyte concentration on agreement is of interest. Complementing the bench evaluation, a second phase of agreement evaluation should use mixed field specimens targeting 50% prevalence (to reduce the agreement by chance) and covering the full spectrum of infection stages. Cross-sectional sampling to represent a defined population would not be required, provided that the specimens realistically represent future sample submissions.

6.3.2 Diagnostic sensitivity (DSe) and specificity (DSp)(Chapter IV)

In the case of diagnostic trueness, DSe and DSp already represent specific estimates of diagnostic trueness within D+ and D-, respectively. In Chapter IV, changing classification patterns among salmon populations tested with 5 different detection methods were explained. The modified latent class model (LCM) identified 3 classes of fish (A, B & C). Although class A and C were identified with certainty as non-infected and infected salmon respectively, the infection status of class B fish could not be clearly

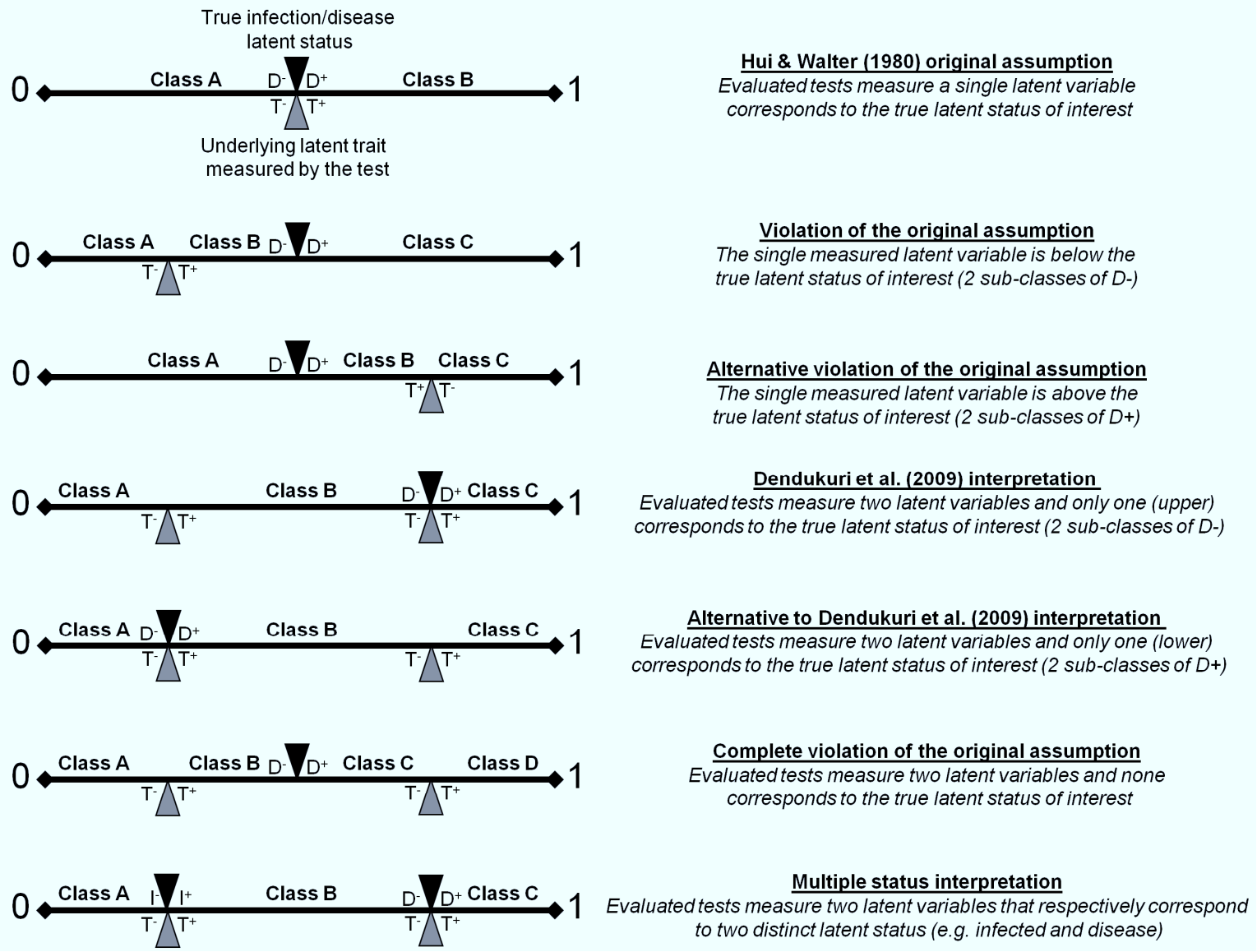


Fig. 6.3. Comparison of relative correspondance between the measured latent variable by the evaluated tests and the true health status. D^+: diseased; D^-: non-diseased; I^+: infected; I^-: non-infected; T^+: positive test result; T^-: negative test result.

identified in this specific case. Multiple class models violate the assumption that all tests involved measure the same underlying latent trait that reflects the true latent health status (Hui & Walter, 1980). When tests detect biological markers that are poorly correlated (e.g. DNA and serum antibody), it is expected that they will not measure the same latent variable. For instance, when two separate latent variables are measured and only one is properly associated with the health status, three classes of animal can be distinguished (Dendukuri et al., 2009). Several hypothetical scenarios can be considered depending on whether one or two latent variables are measured by the tests, and if the latent variable corresponds to the true infection and/or disease status (Fig. 6.3). Each class corresponds to the probability of different test results and further investigation must be conducted to correlate each class with a specific covariate. The estimation of covariate-specific parameters is of most value if prediction of overall test accuracy is based on biologically sound information about the covariate distribution in the population.

6.4 Application and use of covariate-specific estimates

6.4.1 Prediction of test accuracy in external populations

Predicting test performances for an external population can be done by covariate-specific estimate weighting (Björk et al., 2009). As illustrated in Chapter III, the test operating characteristics were predicted for a range of proportions of the main covariate factor of interest (i.e. infection) and of mixture distributions of a secondary factor (i.e. infection stage). Using a modelling approach, agreement was estimated for any possible

population profile giving the full range covered by the test agreement. Although diagnostic trueness is conventionally reported separately for D+ and D- (DSe and DSp, respectively), prediction of DSe and DSp can also be done by weighting sub-categories within D+ or D-. For instance, in Chapter IV, the estimation model identified 3 classes of fish (A, B and C) within which test performances differed. Assuming the identity and the proportion of class B salmon relative to the 2 other classes (A or C), the overall DSe or DSp could be predicted (Fig. 6.4). In this instance, when class B fish are assumed infected, the overall DSe can be computed using a DSe-weighted average based on anticipated proportions of class B and C fish (Fig. 6.4A). The DSe of nucleic acid amplification assays (NAATs) appeared more stable compared to other assays across the mixture distribution of low- and high-infected fish within the infected salmon group. When class B fish were assumed non-infected, virus isolation (VI) and antibody-based assays (ABAs) revealed a much more stable DSp (Fig. 6.4B).

Prediction assumed that test accuracy is constant within each covariate category and does not vary with the degree of representation of other covariates in the same population. For instance, it is assumed that DSp does not vary with the proportion of D+ (prevalence). Various studies have, however, challenged this assumption (e.g. Leeflang et al., 2009). It is reasonable to think that test performance in D- individuals (DSp) is dependent on the pressure of cross-contamination coming from D+ samples. As the proportion of D+ in the sampled population (prevalence) and/or sample pool becomes larger, then the pressure and chance of cross-contamination likely increases. Therefore, DSp and agreement in non-infected individuals may also vary across the range prevalences. The violation of these assumptions may introduce estimation bias for

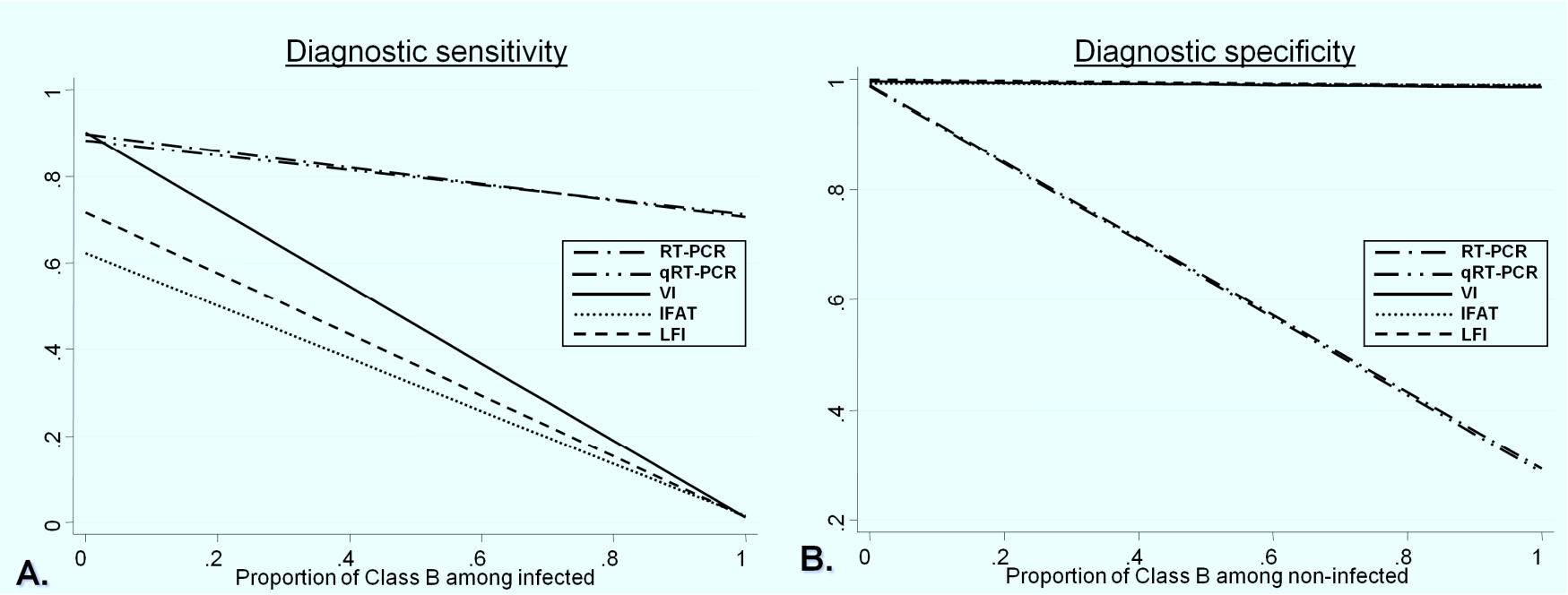


Fig. 6.4. Linear progression of overall diagnostic accuracy.
(A) Overall diagnostic sensitivity (when class B is considered infected); (B) Overall diagnostic specificity (when class B is considered non-infected), with increasing proportion of class B fish among infected (or non-infected) fish for each of the 5 studied tests. Overall estimates were computed as proportion-weighted averages of class-specific estimates. RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-transcriptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFI: lateral flow immunoassay.

predicting test accuracy. For instance, specific agreement predicted for D- fish (0% prevalence) may have been underestimated since agreement estimation within D- fish derived from a sample pool of 48% assumed prevalence corresponding to a much stronger pressure of cross-contamination. This finding supports the assumption that constant performance within covariate groups might not be appropriate and requires further investigation to estimate the degree of dependence between accuracy and covariate proportions (e.g. prevalence). If strong dependence of accuracy on prevalence is modelled, variation of accuracy can be predicted from prevalence (Brenner & Gefeller, 1997). For instance, DSe and DSp could be assessed for a 50% prevalence population and predicted across prevalences using information about population infection distribution (e.g. normal or bimodal). The association between DSe (or DSp) and prevalence can thereafter be directly included in the prediction procedure.

Another application of covariate weighted estimates involves the selection of cutpoints as illustrated in Chapter V. Applied to the special case of real-time RT-PCR (qRT-PCR) for ISAV in salmon, the profile of the two-graph receiver operating characteristic (TG-ROC) curve depends upon the distribution of the influencing covariate factors within either D+ or D- individuals. The TG-ROC can also be predicted from covariate-specific curves by weighted averages, resulting in a substantially changed cutpoint selection.

Nonetheless, to predict test accuracy requires an external population with known or assumed prevalence and spectrum of ISAV loads within the population. Unfortunately, little or no information is yet available about the ISAV infection dynamics and associated clinical information at the fish or population level. Future research should focus on ISAV

dynamics with descriptions of the interdependence of various covariate factors. For instance, salmon cages experiencing a severe episode of ISA with the HPR4 virulent strain were monitored throughout the outbreak. By regular sampling and testing with qRT-PCR to identify their infection status and stage (virus loads). Preliminary results revealed that infected salmon in a population were either highly or lowly infected, while very few intermediate stages were detected. This unexpected observation led to a hypothetical three dimensional model of the disease dynamic during the outbreak (Fig. 6.5). Although, this hypothesis still needs to be verified, a precise estimation of the prevalence and virus load distribution in the population over time will further elucidate cage level disease dynamics during an outbreak, and refine the prediction of test accuracy. A detailed comprehension of the infection dynamics in salmon populations enables the use of covariate-specific estimates to adapt and optimize the testing strategy.

6.4.2 Selection of strategy to fit purpose

Once the overall estimates of test accuracy have been predicted from the covariate-specific estimates and the assumed covariate distribution in the target population, an appropriate testing strategy can be designed to fit the intended purpose (outlines in Chapter I). Depending on the intended purpose, the selection of the test(s) may differ substantially. In general, the primary parameters of interest when selecting diagnostic tests are the predictive values, a concept illustrated by the evaluation estimates from the operating characteristics of the 5 ISAV detection assays in Chapter IV.

For domestic surveillance, the positive predictive value of a positive test result

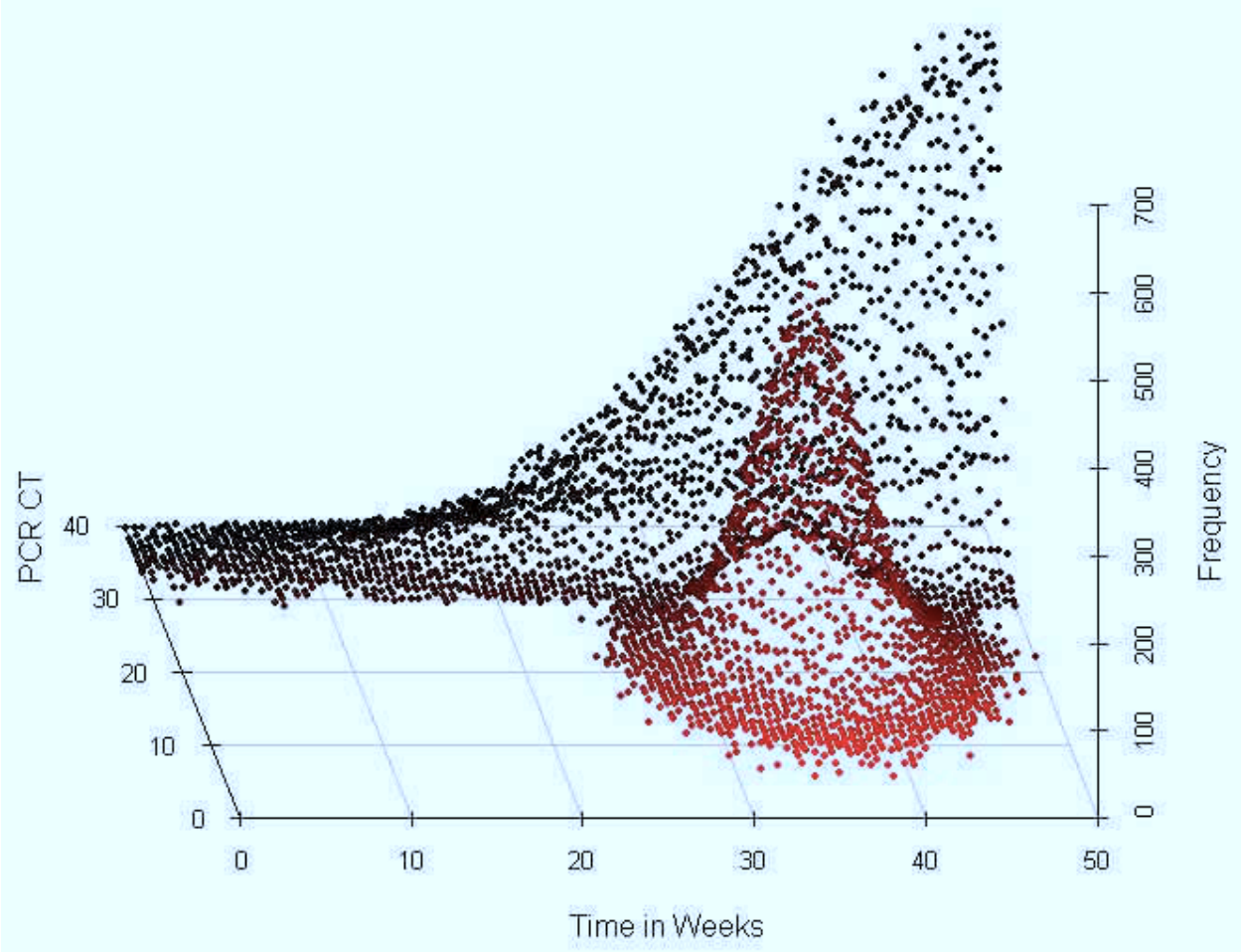
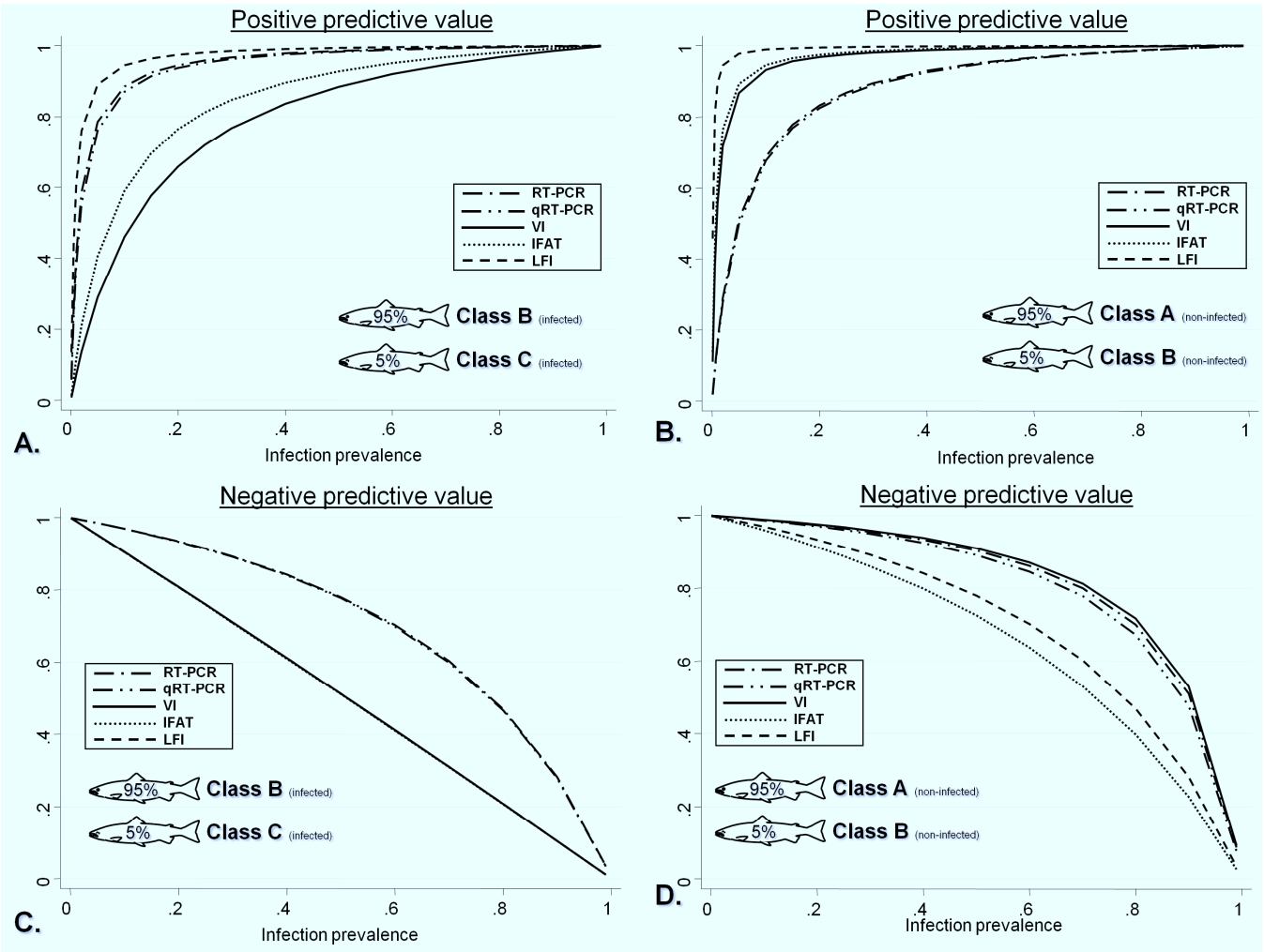


Fig. 6.5. Hypothesized 3D Histogram of an ISAV outbreak.
The frequency of infection is represented as a function of time (weeks) and virus load (low values of PCR CT refer to high virus load, in red, while low values of PCR CT values refer to low virus load, in black). Highly infected fish (in red) are assumed to match with clinical mortality, and low-infected fish are assumed to be apparently healthy fish.

(PPV) is of interest and depends on test operating characteristics (DSe/DSp) and the assumed prevalence in the sampled population. In the situation where class B fish are assumed to be lowly infected salmon, they are expected to be predominant in the infection class during early stages of infection. For each test, the overall DSe can be deduced from Fig. 6.4, assuming a specific ratio of class B and C fish. Thereafter PPV can be estimated for different prevalences using the conventional formula (Dohoo et al., 2009). For instance, PPVs for each of the 5 tests were calculated based on a proportion of 95% of class B and 5% of class C fish across infection prevalences (Fig. 6.6A). With the highest DSp, the lateral flow immunoassay (LFI) would be the best choice to optimize PPV. When only class B fish are present at the start of the infection, the NAATs become the techniques of choice since they alone detected class B fish (data not shown). VI was the worst performing test in both scenarios (Fig. 6.6A). Conversely, if class B fish were assumed to be non-infected salmon, PPV was computed using a DSp-weighted average based on proportions of class A and B fish. However, in early infected populations, recovering fish carrying residual RNA would be rare or absent. PPVs for each of the 5 tests were thus computed for fish proportions of 5 % class B and 95% class A across infection prevalences (Fig. 6.6B). Individually, LFI performed the best and NAATs the worst (i.e. lowest DSp). However, when only class A fish are considered, NAATs clustered with other tests since NAAT's DSp were mainly impacted by testing positive class B fish (data not shown).

For confirmatory purposes, the parameter of interest is the negative predictive value of a negative test result (NPV). Similarly, NPV was computed for both scenarios of class B identities (i.e. infected or not). When class B fish were assumed to be low-



6.6. Predictive value estimates for the 5 tests across infection prevalence. Two scenariosare compared: class B assumed to be infected (A & C), and c
sumed to be non-infected (B & D). Probability to be infected given test positive (positive predictive value) when, among infected fish, 95% are class B ar
class C (A). Probability to be infected given test positive when, among non-infected fish, 95% are class A and 5% class B (B). Probability to be non-infec
n test negative (negative predictive value) when, among infected, fish 95% are class B and 5% class C (C). Probability to be non-infected given test negat
n, among non-infected fish, 95% are class A and 5% class B (D). RT-PCR: reverse-transcriptase polymerase chain reaction; qRT-PCR: real-time reverse-
criptase polymerase chain reaction; VI: virus isolation; IFAT: indirect fluorescent antibody test; LFIA: lateral flow immunoassay.

infected salmon, the NAATs were clearly the methods of choice (Fig. 6.6C), while VI became the technique to use when class B fish were considered to be free from infection (Fig. 6.6D).

Although classification performance might be of primary interest for the interpretation of surveillance programs, economics may be the greatest influence on the selection of test strategies. For instance, the use of multiple tests interpreted in series or parallel, should be considered when optimizing performance and/or reducing cost (Dohoo et al., 2009). Evaluation and comparison of sampling and testing strategies specifically applied to ISAV detection in various stages of Atlantic salmon production was suggested by Nérette et al. (2008) and this provided a good example of diagnostic test estimation application.

Evaluation of diagnostic test performance is of special interest for international trade. According to bilateral agreements, demonstration of freedom from infection/disease might be necessary to exchange animals and animal products between countries. Recently developed, output-based standards to assess surveillance programs were adopted by the OIE to document freedom of disease (OIE, 2009). The evaluation of confidence levels for freedom from infection/disease (probability of freedom) requires the evaluation of the sensitivity of surveillance systems computed using individual estimates of DSe and DSp of tests used (More at al., 2009). Adapted from a decision-tree analysis, the computation can be adapted to several classes of animals associated with different test result probabilities. In a decision tree, we suggest that different classes of infected fish be included as a "detection category node" after the "infection node" at the animal level.

6.5 Conclusions

In conclusion, diagnostic test accuracy depends on many biological factors associated with the manifestation of a disease in the population, as seen with the evaluation of tests for ISAV in salmon in this thesis. The infection status and the severity of infection were the two most obvious factors that influenced the accuracy of the assays. Therefore, specific estimates of test operating characteristics for particular covariates (e.g. infection severity) seem an intuitive approach to properly evaluate diagnostic tests. It was determined that test performance in external populations may be predicted by covariate-specific estimate weighting. However, the application of this approach requires two basic assumptions: i) the diagnostic test accuracy must be constant within each covariate pattern and does not depend on the proportion of other factors in the population, and ii) the dynamics of the infection in the population, including the proportions of infection stages, must be known and progress predictably. When the only accessible populations are free from infection, diagnostic test evaluations are virtually impossible. Knowledgeable descriptions of infection/disease proportions are critical when developing test strategies that fit the purpose in a population.

6.5 References

Björk, J., Grubb, A., Nyman, U., (2009). Variability in diagnostic accuracy can be estimated using simple population weighting. J. Clin. Epidemiol. 62, 54-7.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C.W., 2003. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin. Radiol. 58, 575-580.

Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat. Med. 16, 981-991.

Dendukuri, N., Hadgu, A., Wang, L., 2009. Modeling conditional dependence between diagnostic tests: A multiple latent variable model. Stat. Med. 28:441-461.

Dohoo, I., Martin, W., Stryhn, H. (Eds.), 2009. Veterinary Epidemiologic Research. 2nd ed., AVC Inc., Charlottetown, Canada.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.

Leeflang, M.M., Bossuyt, P.M., Irwig, L., 2009. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J. Clin. Epidemiol. 62, 5-12.

More, S.J., Cameron, A.R., Greiner, M., Clifton-Hadley, R.S., Rodeia, S.C., Bakker, D., Salman, M.D., Sharp, J.M., De Massis, F., Aranaz, A., Boniotti, M.B., Gaffuri, A., Have, P., Verloo, D., Woodford, M., Wierup, M., 2009. Defining output-based standards to achieve and maintain tuberculosis freedom in farmed deer, with reference to member states of the European Union. Prev. Vet. Med. 90, 254-67.

Nérette, P., Hammell, L., Dohoo, I., Gardner I., 2008. Evaluation of testing strategies for infectious salmon anaemia and implications for surveillance and control programs. Aquaculture 280, 53-59.

Office International des Epizooties, 2009. Manual of Diagnostic Tests for Aquatic Animals 2009. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 10-30.

Yerushalmy, J., 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Publ. Health Rep. 62, 1432-1449.

APPENDIX 1

Likelihood ratio of a positive test result (LR^+) as function of diagnostic sensitivity (DSe) and specificity (DSp). LR^+ was computed as *DSe / (1-DSp)*.

Estimate of test diagnostic sensitivity (%)	Estimate of the test diagnostic specificity (%)																				
LR^+	**90**	**90.5**	**91**	**91.5**	**92**	**92.5**	**93**	**93.5**	**94**	**94.5**	**95**	**95.5**	**96**	**96.5**	**97**	**97.5**	**98**	**98.5**	**99**	**99.5**	**99.9**
99.9	10.0	10.5	11.1	11.8	12.5	13.3	14.3	15.4	16.7	18.2	20.0	22.2	25.0	28.5	33.3	40.0	50.0	66.6	100	200	999
99	9.9	10.4	11.0	11.6	12.4	13.2	14.1	15.2	16.5	18.0	19.8	22.0	24.8	28.3	33.0	39.6	49.5	66.0	99.0	198	990
98	9.8	10.3	10.9	11.5	12.3	13.1	14.0	15.1	16.3	17.8	19.6	21.8	24.5	28.0	32.7	39.2	49.0	65.3	98.0	196	980
97	9.7	10.2	10.8	11.4	12.1	12.9	13.9	14.9	16.2	17.6	19.4	21.6	24.3	27.7	32.3	38.8	48.5	64.7	97.0	194	970
96	9.6	10.1	10.7	11.3	12.0	12.8	13.7	14.8	16.0	17.5	19.2	21.3	24.0	27.4	32.0	38.4	48.0	64.0	96.0	192	960
95	9.5	10.0	10.6	11.2	11.9	12.7	13.6	14.6	15.8	17.3	19.0	21.1	23.8	27.1	31.7	38.0	47.5	63.3	95.0	190	950
94	9.4	9.9	10.4	11.1	11.8	12.5	13.4	14.5	15.7	17.1	18.8	20.9	23.5	26.9	31.3	37.6	47.0	62.7	94.0	188	940
93	9.3	9.8	10.3	10.9	11.6	12.4	13.3	14.3	15.5	16.9	18.6	20.7	23.3	26.6	31.0	37.2	46.5	62.0	93.0	186	930
92	9.2	9.7	10.2	10.8	11.5	12.3	13.1	14.2	15.3	16.7	18.4	20.4	23.0	26.3	30.7	36.8	46.0	61.3	92.0	184	920
91	9.1	9.6	10.1	10.7	11.4	12.1	13.0	14.0	15.2	16.5	18.2	20.2	22.8	26.0	30.3	36.4	45.5	60.7	91.0	182	910
90	9.0	9.5	10.0	10.6	11.3	12.0	12.9	13.8	15.0	16.4	18.0	20.0	22.5	25.7	30.0	36.0	45.0	60.0	90.0	180	900
89	8.9	9.4	9.9	10.5	11.1	11.9	12.7	13.7	14.8	16.2	17.8	19.8	22.3	25.4	29.7	35.6	44.5	59.3	89.0	178	890
88	8.8	9.3	9.8	10.4	11.0	11.7	12.6	13.5	14.7	16.0	17.6	19.6	22.0	25.1	29.3	35.2	44.0	58.7	88.0	176	880
87	8.7	9.2	9.7	10.2	10.9	11.6	12.4	13.4	14.5	15.8	17.4	19.3	21.8	24.9	29.0	34.8	43.5	58.0	87.0	174	870
86	8.6	9.1	9.6	10.1	10.8	11.5	12.3	13.2	14.3	15.6	17.2	19.1	21.5	24.6	28.7	34.4	43.0	57.3	86.0	172	860
85	8.5	8.9	9.4	10.0	10.6	11.3	12.1	13.1	14.2	15.5	17.0	18.9	21.3	24.3	28.3	34.0	42.5	56.7	85.0	170	850
84	8.4	8.8	9.3	9.9	10.5	11.2	12.0	12.9	14.0	15.3	16.8	18.7	21.0	24.0	28.0	33.6	42.0	56.0	84.0	168	840
83	8.3	8.7	9.2	9.8	10.4	11.1	11.9	12.8	13.8	15.1	16.6	18.4	20.8	23.7	27.7	33.2	41.5	55.3	83.0	166	830
82	8.2	8.6	9.1	9.6	10.3	10.9	11.7	12.6	13.7	14.9	16.4	18.2	20.5	23.4	27.3	32.8	41.0	54.7	82.0	164	820
81	8.1	8.5	9.0	9.5	10.1	10.8	11.6	12.5	13.5	14.7	16.2	18.0	20.3	23.1	27.0	32.4	40.5	54.0	81.0	162	810
80	8.0	8.4	8.9	9.4	10.0	10.7	11.4	12.3	13.3	14.5	16.0	17.8	20.0	22.9	26.7	32.0	40.0	53.3	80.0	160	800
79	7.9	8.3	8.8	9.3	9.9	10.5	11.3	12.2	13.2	14.4	15.8	17.6	19.8	22.6	26.3	31.6	39.5	52.7	79.0	158	790
78	7.8	8.2	8.7	9.2	9.8	10.4	11.1	12.0	13.0	14.2	15.6	17.3	19.5	22.3	26.0	31.2	39.0	52.0	78.0	156	780
77	7.7	8.1	8.6	9.1	9.6	10.3	11.0	11.8	12.8	14.0	15.4	17.1	19.3	22.0	25.7	30.8	38.5	51.3	77.0	154	770
76	7.6	8.0	8.4	8.9	9.5	10.1	10.9	11.7	12.7	13.8	15.2	16.9	19.0	21.7	25.3	30.4	38.0	50.7	76.0	152	760
75	7.5	7.9	8.3	8.8	9.4	10.0	10.7	11.5	12.5	13.6	15.0	16.7	18.8	21.4	25.0	30.0	37.5	50.0	75.0	150	750
74	7.4	7.8	8.2	8.7	9.3	9.9	10.6	11.4	12.3	13.5	14.8	16.4	18.5	21.1	24.7	29.6	37.0	49.3	74.0	148	740
73	7.3	7.7	8.1	8.6	9.1	9.7	10.4	11.2	12.2	13.3	14.6	16.2	18.3	20.9	24.3	29.2	36.5	48.7	73.0	146	730
72	7.2	7.6	8.0	8.5	9.0	9.6	10.3	11.1	12.0	13.1	14.4	16.0	18.0	20.6	24.0	28.8	36.0	48.0	72.0	144	720
71	7.1	7.5	7.9	8.4	8.9	9.5	10.1	10.9	11.8	12.9	14.2	15.8	17.8	20.3	23.7	28.4	35.5	47.3	71.0	142	710
70	7.0	7.4	7.8	8.2	8.8	9.3	10.0	10.8	11.7	12.7	14.0	15.6	17.5	20.0	23.3	28.0	35.0	46.7	70.0	140	700
69	6.9	7.3	7.7	8.1	8.6	9.2	9.9	10.6	11.5	12.5	13.8	15.3	17.3	19.7	23.0	27.6	34.5	46.0	69.0	138	690
68	6.8	7.2	7.6	8.0	8.5	9.1	9.7	10.5	11.3	12.4	13.6	15.1	17.0	19.4	22.7	27.2	34.0	45.3	68.0	136	680
67	6.7	7.1	7.4	7.9	8.4	8.9	9.6	10.3	11.2	12.2	13.4	14.9	16.8	19.1	22.3	26.8	33.5	44.7	67.0	134	670
66	6.6	6.9	7.3	7.8	8.3	8.8	9.4	10.2	11.0	12.0	13.2	14.7	16.5	18.9	22.0	26.4	33.0	44.0	66.0	132	660
65	6.5	6.8	7.2	7.6	8.1	8.7	9.3	10.0	10.8	11.8	13.0	14.4	16.3	18.6	21.7	26.0	32.5	43.3	65.0	130	650
64	6.4	6.7	7.1	7.5	8.0	8.5	9.1	9.8	10.7	11.6	12.8	14.2	16.0	18.3	21.3	25.6	32.0	42.7	64.0	128	640
63	6.3	6.6	7.0	7.4	7.9	8.4	9.0	9.7	10.5	11.5	12.6	14.0	15.8	18.0	21.0	25.2	31.5	42.0	63.0	126	630
62	6.2	6.5	6.9	7.3	7.8	8.3	8.9	9.5	10.3	11.3	12.4	13.8	15.5	17.7	20.7	24.8	31.0	41.3	62.0	124	620
61	6.1	6.4	6.8	7.2	7.6	8.1	8.7	9.4	10.2	11.1	12.2	13.6	15.3	17.4	20.3	24.4	30.5	40.7	61.0	122	610
60	6.0	6.3	6.7	7.1	7.5	8.0	8.6	9.2	10.0	10.9	12.0	13.3	15.0	17.1	20.0	24.0	30.0	40.0	60.0	120	600

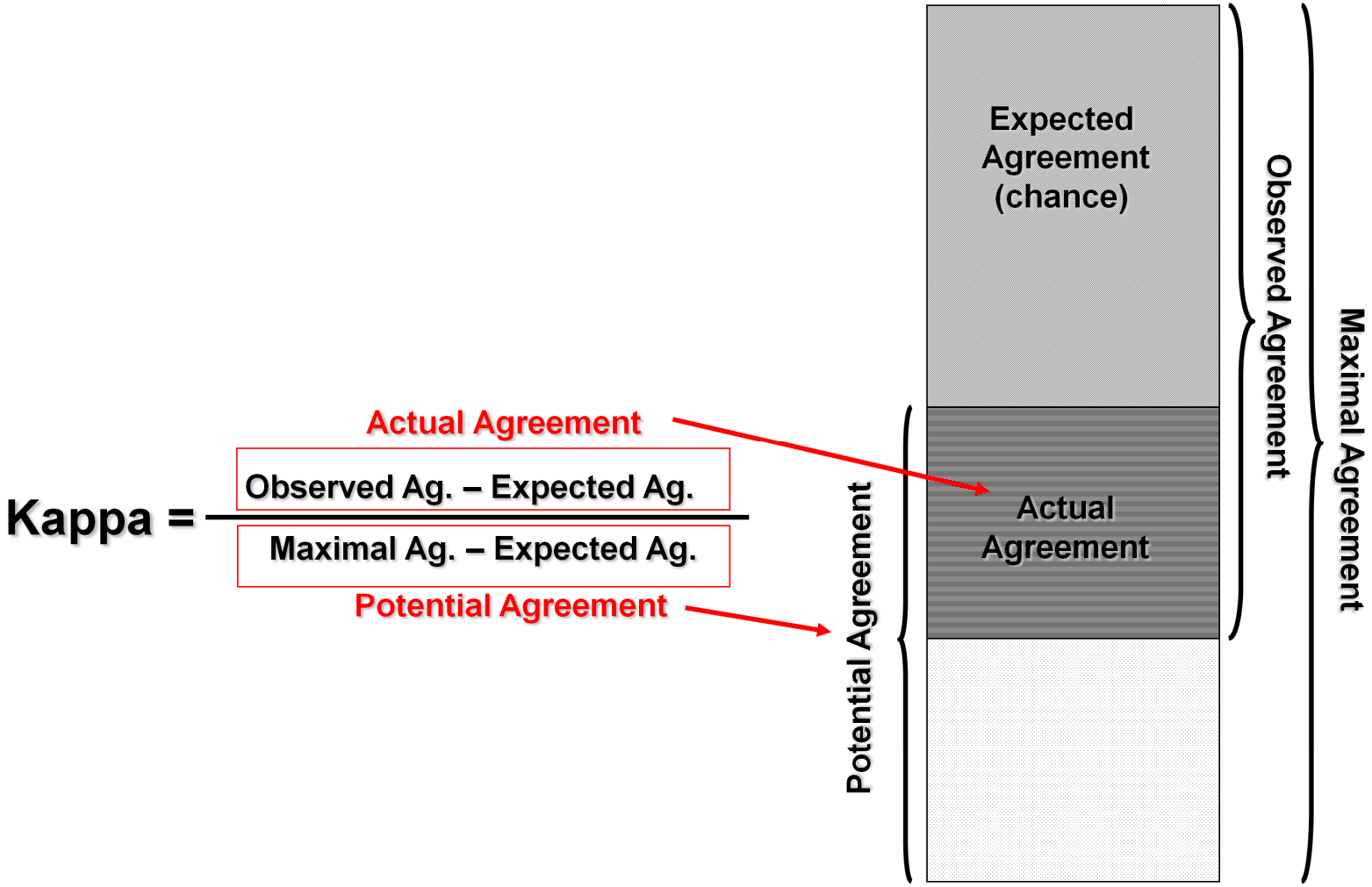
APPENDIX 2

Likelihood ratio of a negative test result (LR^-) as function of diagnostic sensitivity (DSe) and specificity (DSp). *LR^-* was computed as *(1-DSe) / DSp.*

Estimate of test diagnostic sensitivity (%)	Estimate of the test diagnostic specificity (%)																				
LR^-	**90**	**90.5**	**91**	**91.5**	**92**	**92.5**	**93**	**93.5**	**94**	**94.5**	**95**	**95.5**	**96**	**96.5**	**97**	**97.5**	**98**	**98.5**	**99**	**99.5**	**99.9**
99.9	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
99	.011	.011	.011	.011	.011	.011	.011	.011	.011	.011	.011	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010
98	.022	.022	.022	.022	.022	.022	.022	.021	.021	.021	.021	.021	.021	.021	.021	.021	.020	.020	.020	.020	.020
97	.033	.033	.033	.033	.033	.032	.032	.032	.032	.032	.032	.031	.031	.031	.031	.031	.031	.030	.030	.030	.030
96	.044	.044	.044	.044	.043	.043	.043	.043	.043	.042	.042	.042	.042	.041	.041	.041	.041	.041	.040	.040	.040
95	.056	.055	.055	.055	.054	.054	.054	.053	.053	.053	.053	.052	.052	.052	.052	.051	.051	.051	.051	.050	.050
94	.067	.066	.066	.066	.065	.065	.065	.064	.064	.063	.063	.063	.063	.062	.062	.062	.061	.061	.061	.060	.060
93	.078	.077	.077	.077	.076	.076	.075	.075	.074	.074	.074	.073	.073	.073	.072	.072	.071	.071	.071	.070	.070
92	.089	.088	.088	.087	.087	.086	.086	.086	.085	.085	.084	.084	.083	.083	.082	.082	.082	.081	.081	.080	.080
91	.100	.099	.099	.098	.098	.097	.097	.096	.096	.095	.095	.094	.094	.093	.093	.092	.092	.091	.091	.090	.090
90	.111	.110	.110	.109	.109	.108	.108	.107	.106	.106	.105	.105	.104	.104	.103	.103	.102	.102	.101	.101	.100
89	.122	.122	.121	.120	.120	.119	.118	.118	.117	.116	.116	.115	.115	.114	.113	.113	.112	.112	.111	.111	.110
88	.133	.133	.132	.131	.130	.130	.129	.128	.128	.127	.126	.126	.125	.124	.124	.123	.122	.122	.121	.121	.120
87	.144	.144	.143	.142	.141	.141	.140	.139	.138	.138	.137	.136	.135	.135	.134	.133	.133	.132	.131	.131	.130
86	.156	.155	.154	.153	.152	.151	.151	.150	.149	.148	.147	.147	.146	.145	.144	.144	.143	.142	.141	.141	.140
85	.167	.166	.165	.164	.163	.162	.161	.160	.160	.159	.158	.157	.156	.155	.155	.154	.153	.152	.152	.151	.150
84	.178	.177	.176	.175	.174	.173	.172	.171	.170	.169	.168	.168	.167	.166	.165	.164	.163	.162	.162	.161	.160
83	.189	.188	.187	.186	.185	.184	.183	.182	.181	.180	.179	.178	.177	.176	.175	.174	.173	.173	.172	.171	.170
82	.200	.199	.198	.197	.196	.195	.194	.193	.191	.190	.189	.188	.188	.187	.186	.185	.184	.183	.182	.181	.180
81	.211	.210	.209	.208	.207	.205	.204	.203	.202	.201	.200	.199	.198	.197	.196	.195	.194	.193	.192	.191	.190
80	.222	.221	.220	.219	.217	.216	.215	.214	.213	.212	.211	.209	.208	.207	.206	.205	.204	.203	.202	.201	.200
79	.233	.232	.231	.230	.228	.227	.226	.225	.223	.222	.221	.220	.219	.218	.216	.215	.214	.213	.212	.211	.210
78	.244	.243	.242	.240	.239	.238	.237	.235	.234	.233	.232	.230	.229	.228	.227	.226	.224	.223	.222	.221	.220
77	.256	.254	.253	.251	.250	.249	.247	.246	.245	.243	.242	.241	.240	.238	.237	.236	.235	.234	.232	.231	.230
76	.267	.265	.264	.262	.261	.259	.258	.257	.255	.254	.253	.251	.250	.249	.247	.246	.245	.244	.242	.241	.240
75	.278	.276	.275	.273	.272	.270	.269	.267	.266	.265	.263	.262	.260	.259	.258	.256	.255	.254	.253	.251	.250
74	.289	.287	.286	.284	.283	.281	.280	.278	.277	.275	.274	.272	.271	.269	.268	.267	.265	.264	.263	.261	.260
73	.300	.298	.297	.295	.293	.292	.290	.289	.287	.286	.284	.283	.281	.280	.278	.277	.276	.274	.273	.271	.270
72	.311	.309	.308	.306	.304	.303	.301	.299	.298	.296	.295	.293	.292	.290	.289	.287	.286	.284	.283	.281	.280
71	.322	.320	.319	.317	.315	.314	.312	.310	.309	.307	.305	.304	.302	.301	.299	.297	.296	.294	.293	.291	.290
70	.333	.331	.330	.328	.326	.324	.323	.321	.319	.317	.316	.314	.313	.311	.309	.308	.306	.305	.303	.302	.300
69	.344	.343	.341	.339	.337	.335	.333	.332	.330	.328	.326	.325	.323	.321	.320	.318	.316	.315	.313	.312	.310
68	.356	.354	.352	.350	.348	.346	.344	.342	.340	.339	.337	.335	.333	.332	.330	.328	.327	.325	.323	.322	.320
67	.367	.365	.363	.361	.359	.357	.355	.353	.351	.349	.347	.346	.344	.342	.340	.338	.337	.335	.333	.332	.330
66	.378	.376	.374	.372	.370	.368	.366	.364	.362	.360	.358	.356	.354	.352	.351	.349	.347	.345	.343	.342	.340
65	.389	.387	.385	.383	.380	.378	.376	.374	.372	.370	.368	.366	.365	.363	.361	.359	.357	.355	.354	.352	.350
64	.400	.398	.396	.393	.391	.389	.387	.385	.383	.381	.379	.377	.375	.373	.371	.369	.367	.365	.364	.362	.360
63	.411	.409	.407	.404	.402	.400	.398	.396	.394	.392	.389	.387	.385	.383	.381	.379	.378	.376	.374	.372	.370
62	.422	.420	.418	.415	.413	.411	.409	.406	.404	.402	.400	.398	.396	.394	.392	.390	.388	.386	.384	.382	.380
61	.433	.431	.429	.426	.424	.422	.419	.417	.415	.413	.411	.408	.406	.404	.402	.400	.398	.396	.394	.392	.390
60	.444	.442	.440	.437	.435	.432	.430	.428	.426	.423	.421	.419	.417	.415	.412	.410	.408	.406	.404	.402	.400

APPENDIX 3

Component of the computation for the Cohen's Kappa coefficient.



APPENDIX 4

Sample size required to estimate diagnostic sensitivity (or specificity) when the true status is known (adapted from Flahaut et al., 2005). For an expected DSe (or DSp), the table gives the number of infected/diseased samples required to estimate with a probability of 95% (1 – β) that DSe is at least higher than a defined value DSe_{min} (i.e. lower bound of the confidence interval) with a minimum confidence level of 95% (one sided, 1 - α).

Expected DSe (or DSp)	DSe_{min}: Minimal acceptable lower confidence limit																				
	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98
0.6	265	1055																			
0.65	115	257	1012																		
0.7	62	110	243	947																	
0.75	38	59	102	224	860																
0.8	24	35	54	92	199	751															
0.85	16	22	31	47	80	169	620														
0.9	11	14	19	26	39	65	133	467	708	1217	2642	10164									
0.91	10	13	17	23	34	55	105	311	434	655	1123	2428	9297								
0.92	9	12	15	21	30	46	85	218	287	400	601	1026	2208	8409							
0.93	8	11	14	18	26	40	69	158	200	263	364	546	927	1982	7499						
0.94	8	10	12	16	23	34	56	118	144	182	238	328	489	824	1751	6566					
0.95	7	9	11	15	20	29	46	89	107	130	163	212	290	430	720	1514	5612				
0.96	6	8	10	13	17	24	38	68	80	95	115	143	185	252	369	612	1271	4635			
0.97	6	7	9	11	15	20	30	52	60	70	82	99	122	157	211	307	501	1021	3635		
0.98	5	6	7	9	12	17	24	40	45	51	58	69	82	100	127	169	241	385	763	2610	
0.99	4	5	6	8	10	13	19	29	32	36	40	46	53	63	76	95	123	170	262	493	1552

In the appropriate settings, the study population contains a mix of non- and infected/diseased individuals. Assuming the studied population prevalence (Pr), the required number of non-infected/non-diseased can thereafter be computed as follows: $n_{NI} = n_I * (1-Pr) / Pr$

Exemple: when DSe is expected to be 0.9 with in a minimum DSe_{min} of 0.75, the sample size requires 65 infected/diseased individuals and 260 non-infected/non-diseased if the assumed prevalence is 20% in the population.

APPENDIX 5

Latent Class Model mechanic for 2 tests and 2 populations

Contingency table of the 2 tests results for Population 1

Population k =1	*Test2=1*	*Test2=0*
Test1=1	P_{111}	P_{101}
Test1=0	P_{111}	P_{101}

Contingency table of the 2 tests results for Population 1

Population k =2	*Test2=1*	*Test2=0*
Test1=1	P_{112}	P_{102}
Test1=0	P_{112}	P_{102}

The degree of freedom for each table is 3 since it can be deduce from 3 cells probability the fourth one. Therefore, for each population, only 3 essential information are provided by the data.

Let's defined P_{ijk} the observed probability of a Test 1 result (i) and a Test 2 result (j) in the population k:

$\rightarrow P_{ijk} = Prob(i,j \mid Pop_k)$

An individual from the population k that yield a test result combination (i,j) and can either be diseased (D^+) or not (D^-), therefore:

$\leftrightarrow P_{ijk} = Prob(i,j \cap D^+ \mid Pop_k) + Prob(i,j \cap D^- \mid Pop_k)$

According to the Bayes theorem:

$\leftrightarrow P_{ijk} = Prob(D^+, Pop_k)\ Prob(i,j \mid D^+, Pop_k) + Prob(D^-, Pop_k)\ Prob(i,j \mid D^-, Pop_k)$

Where $Prob(D^+, Pop_k) = Pr_k$ (prevalence in population k) and $Prob(D^-, Pop_k) = 1 - Pr_k$

$\leftrightarrow P_{ijk} = Pr_k\ Prob(i,j \mid D^+, Pop_k) + (1-Pr_k)\ Prob(i,j \mid D^-, Pop_k)$

Assuming the probability of testing i in Test1 is independent of the probability of testing j in Test2 conditional on the disease status (**1st assumption**):

$\rightarrow P_{ijk} = Pr_k\ Prob(i \mid D^+, Pop_k)\ Prob(j \mid D^+, Pop_k) + (1-Pr_k)\ Prob(i \mid D^-, Pop_k)\ Prob(j \mid D^-, Pop_k)$

For *i* (or *j*) = 1 (positive), Prob(1|D^+, Pop_k) = DSe_k and Prob(0|D^+, Pop_k) = (1-DSe_k)
For *i* (or *j*) = 0 (negative), Prob(1|D^+, Pop_k) = (1-DSp_k) and Prob(0|D^+, Pop_k) = DSp_k

Therefore, we can express P for 4 combinations of the two tests results:

$\rightarrow P_{11k} = DSe_{1,k}\, DSe_{2,k}\, Pr_k + (1\text{-}DSe_{1,k})(1\text{-}DSe_{2,k})\, (1\text{-}Pr_k)$
$\rightarrow P_{10k} = DSe_{1,k}\, (1\text{-}DSe_{2,k}) Pr_k + (1\text{-}DSe_{1,k})(1\text{-}DSe_{2,k})\, (1\text{-}Pr_k)$
$\rightarrow P_{01k} = DSe_{1,k}\, DSe_{2,k}\, Pr_k + (1\text{-}DSe_{1,k})(1\text{-}DSe_{2,k})\, (1\text{-}Pr_k)$
$\rightarrow P_{00k} = DSe_{1,k}\, DSe_{2,k}\, Pr_k + (1\text{-}DSe_{1,k})(1\text{-}DSe_{2,k})\, (1\text{-}Pr_k)$

For a single population *k*, we need to estimate 5 parameters (i.e. DSe & DSp of each test & the prevalence) from 3 equations. As previously mentioned, only 3 of the observed proportions are useful since the fourth can be deduced from the 3 others (free of dependence). Therefore, the dataset degree of freedom is smaller than the number of parameter and the model is not identifiable.

If a second population is added, the degree of freedom doubles (6) and equals the number of parameter only increased by one (i.e. different prevalence in the second population). Therefore, it is required at least 2 populations and 2 tests for the model to be, in theory, identifiable (**2nd assumption**). Furthermore, it is also essential to assume that DSe and DSp of both tests are constant across the two populations (**3rd assumption**).

In summary, basic LCM mechanism requires 3 assumptions:

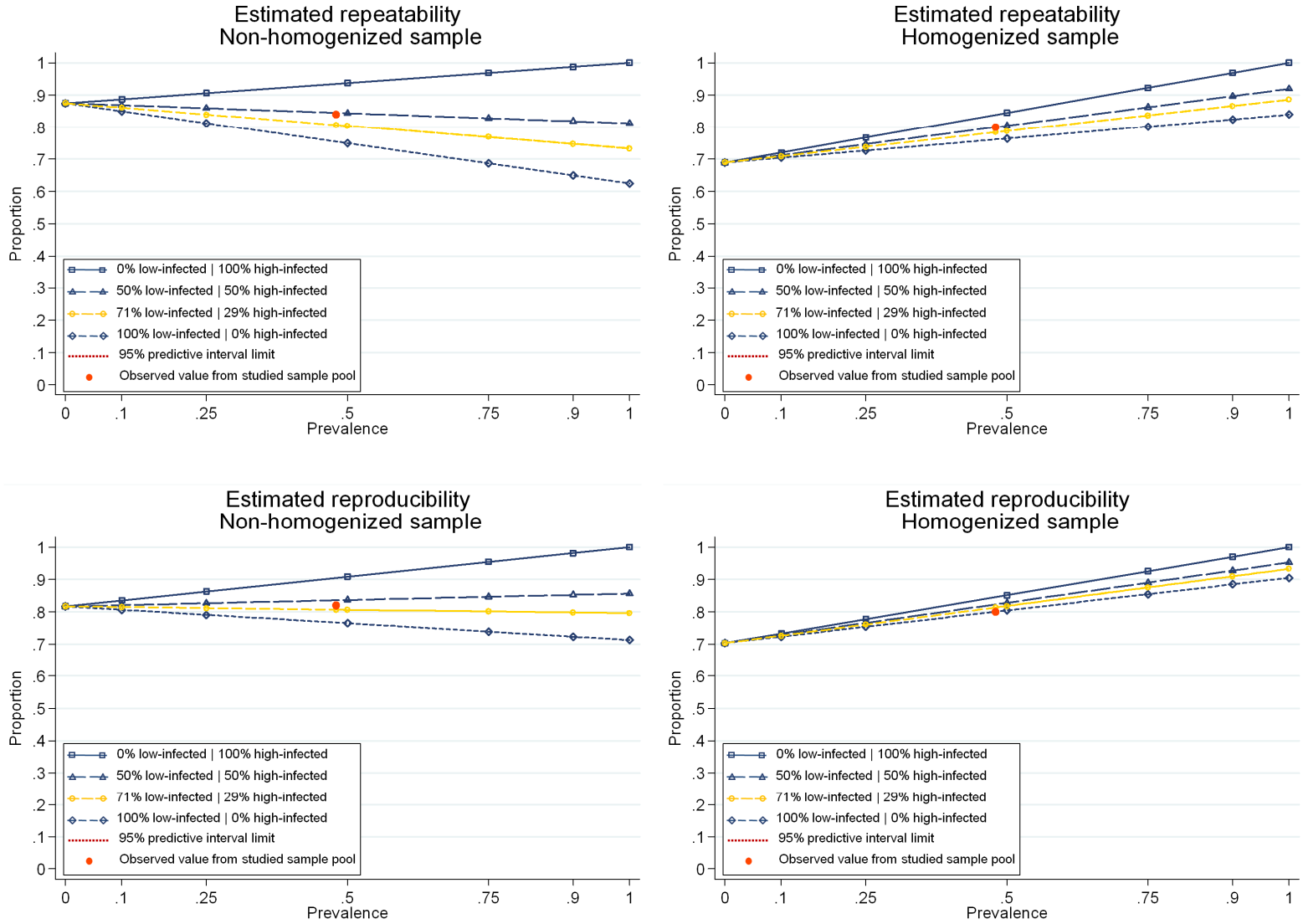
(i) Tests are independent conditional on the disease status

(ii) A minimum of 2 tests run on the same samples from at least 2 populations of different prevalences

(iii) Test operating characteristics (DSE/DSp) are constant across populations

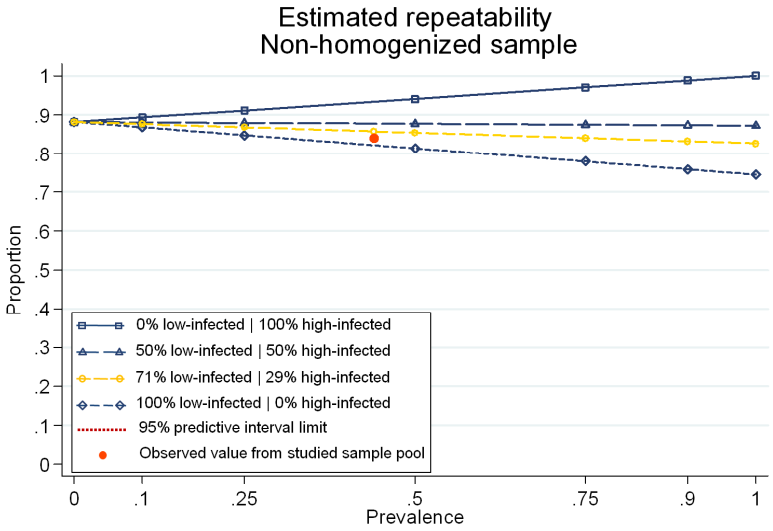
APPENDIX 6

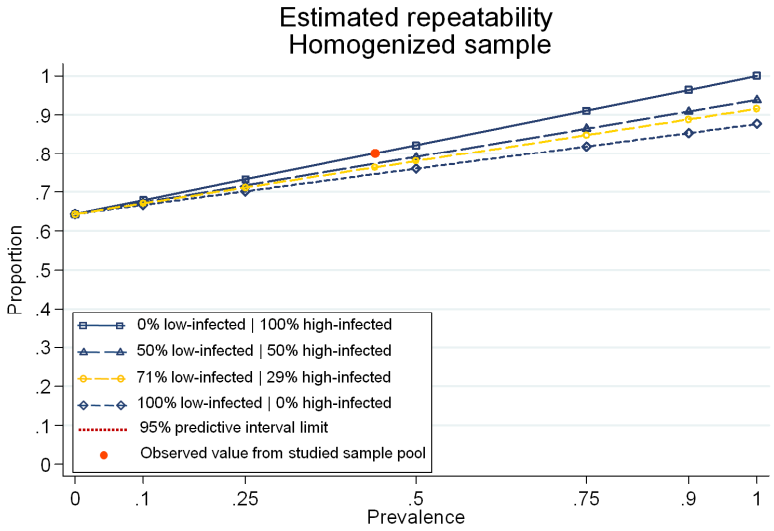
Using PGS definition for fish classification, these graphs predict descriptive estimated repeatability of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and reproducibility for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.

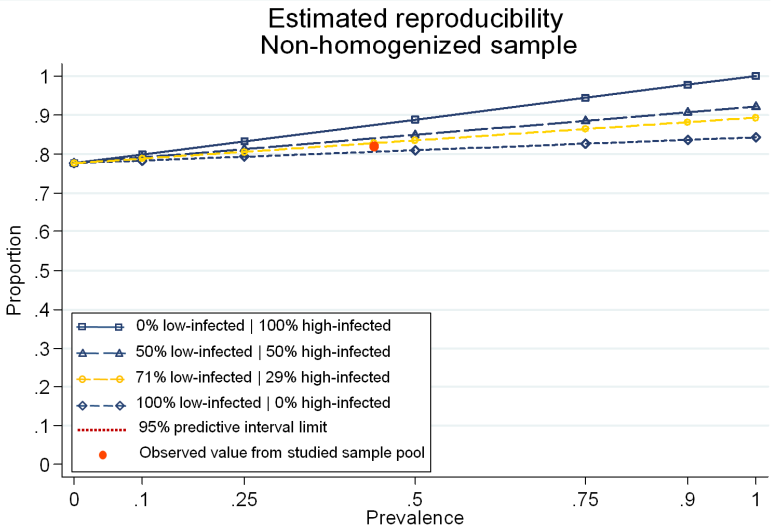


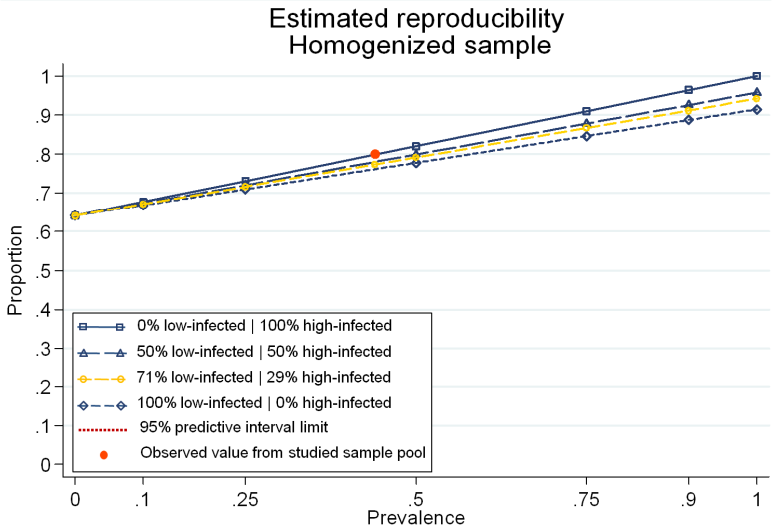
APPENDIX 7

Using Strict-PGS definition for fish classification, these graphs predict descriptive estimated repeatability of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and reproducibility for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.



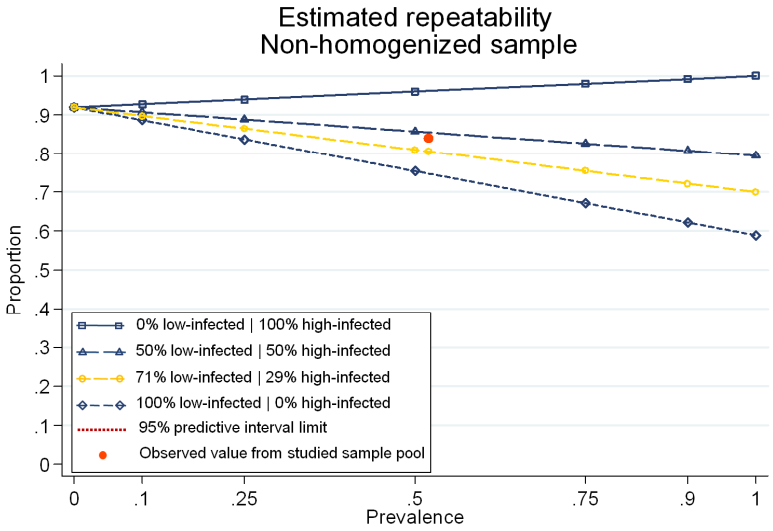


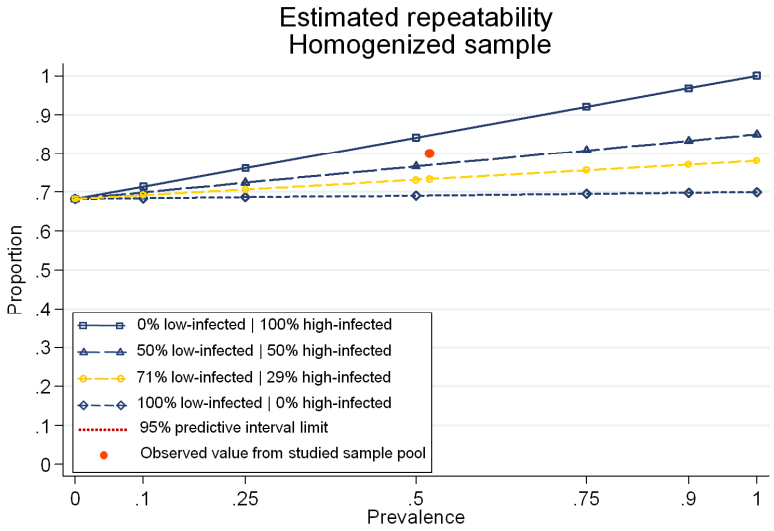


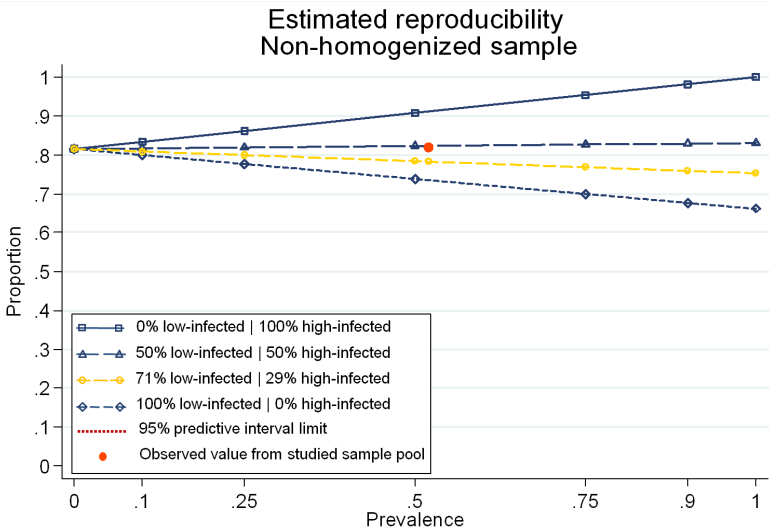


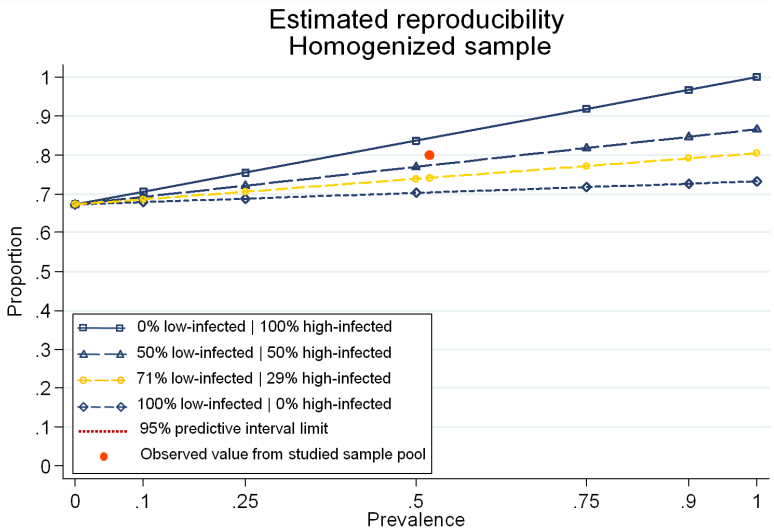
APPENDIX 8

Using Lenient-PGS definition for fish classification, these graphs predict descriptive estimated repeatability of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and reproducibility for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.



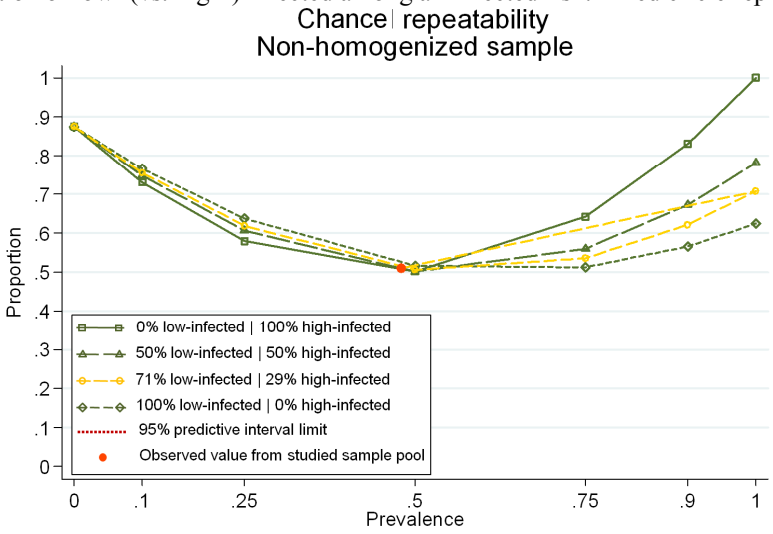


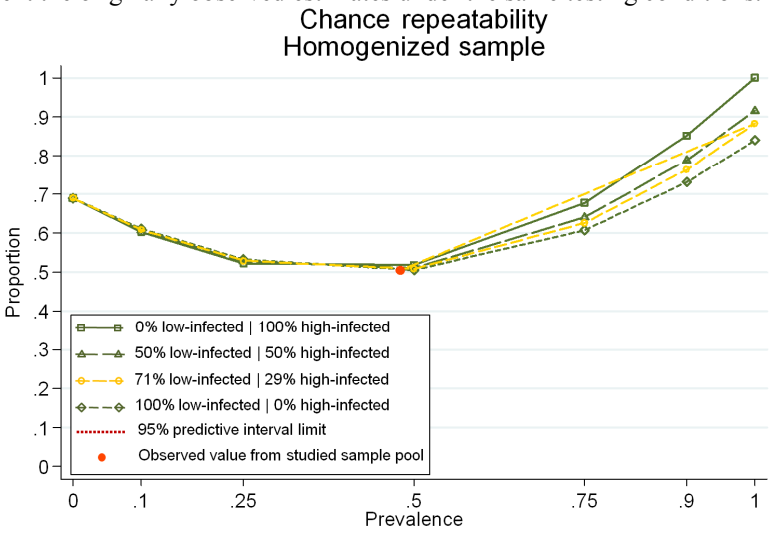


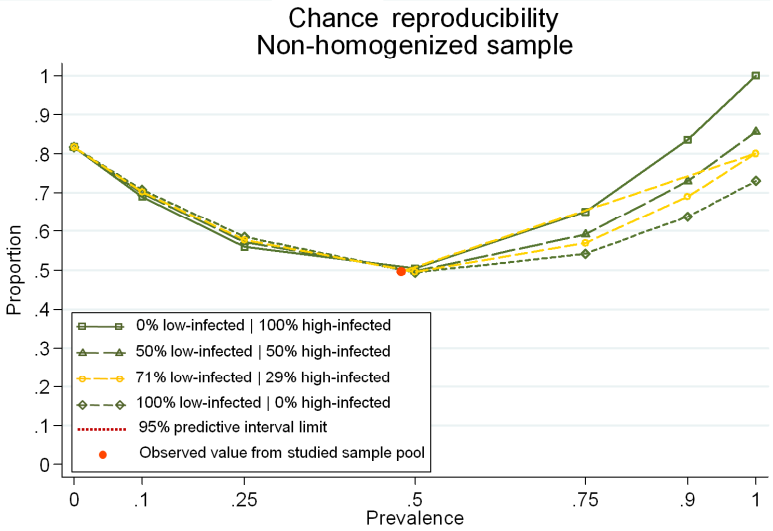


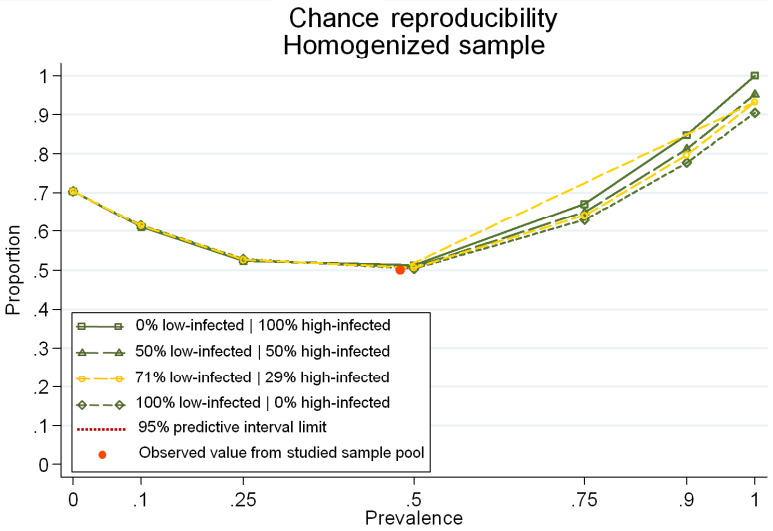
APPENDIX 9

Using PGS definition for fish classification, these graphs predict descriptive chance repeatability of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and reproducibility for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.



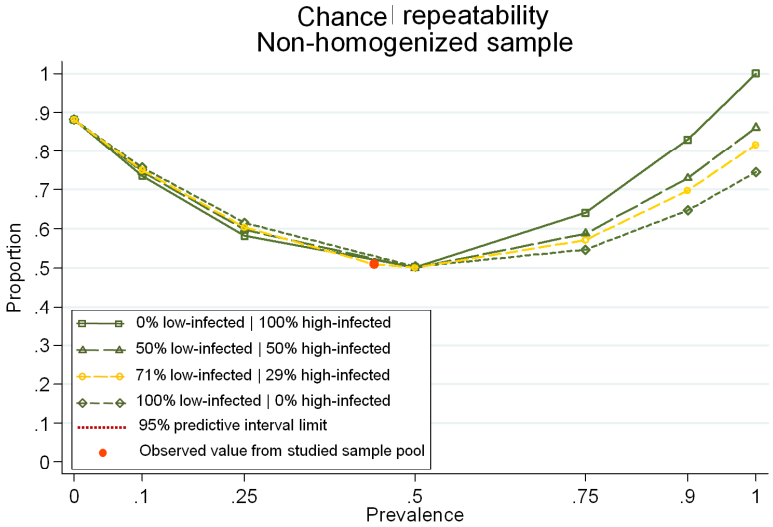


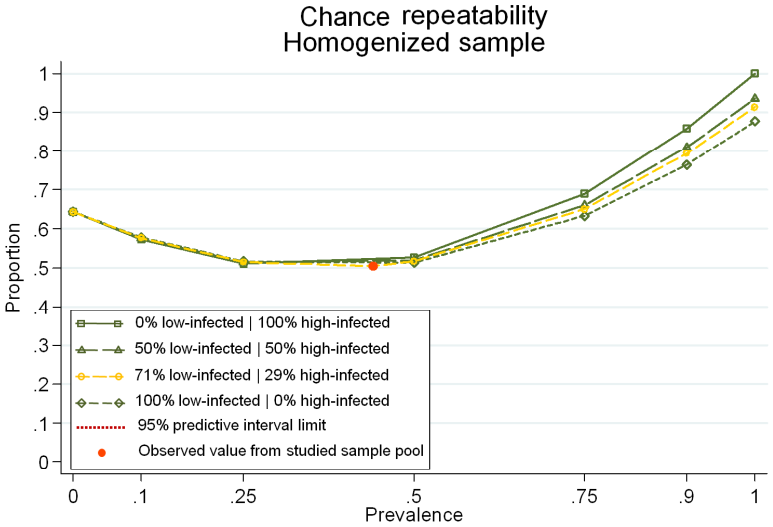


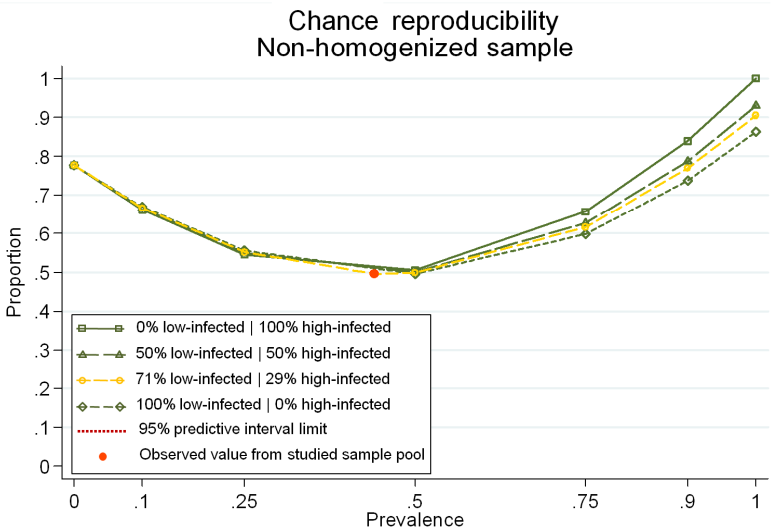


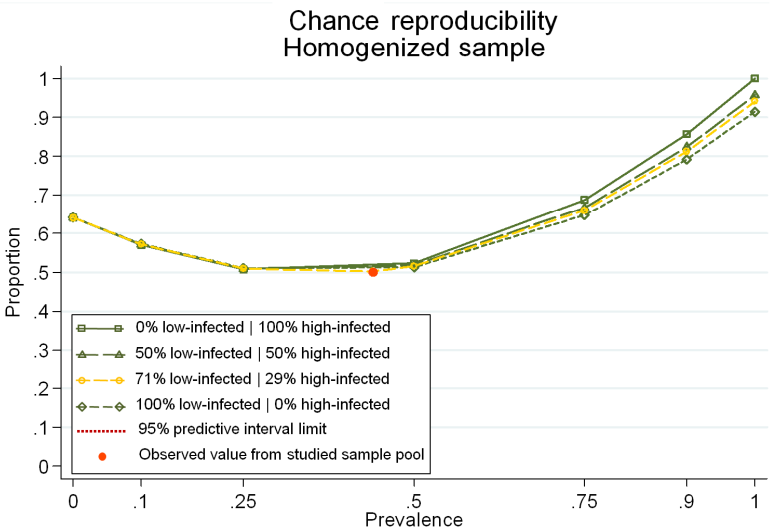
APPENDIX 10

Using Strict-PGS definition for fish classification, these graphs predict descriptive chance repeatability of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and reproducibility for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.



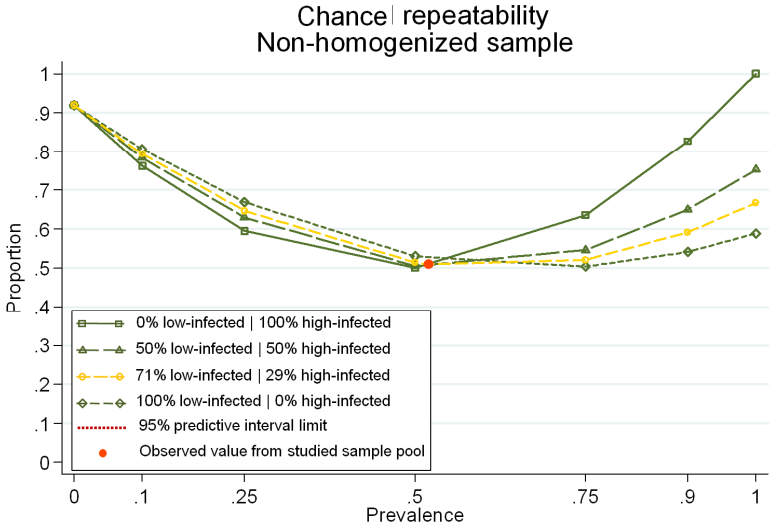


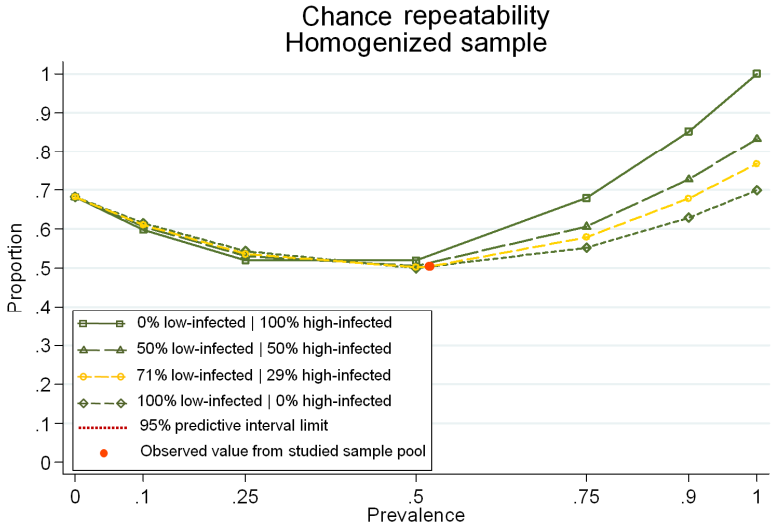


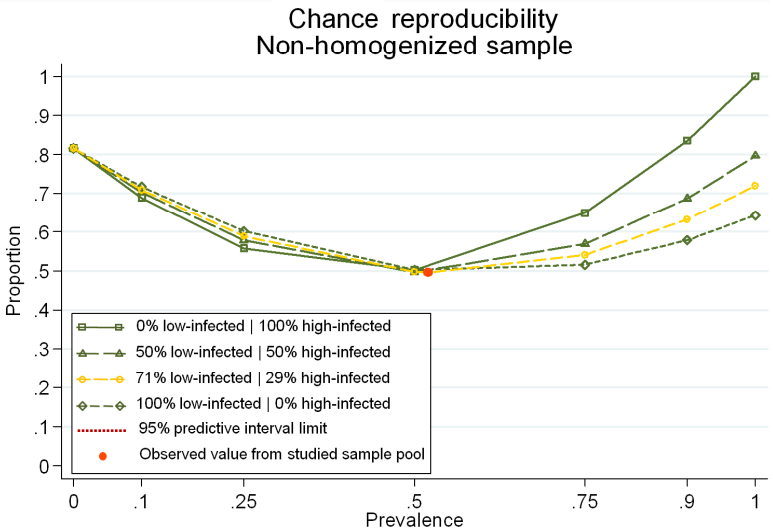


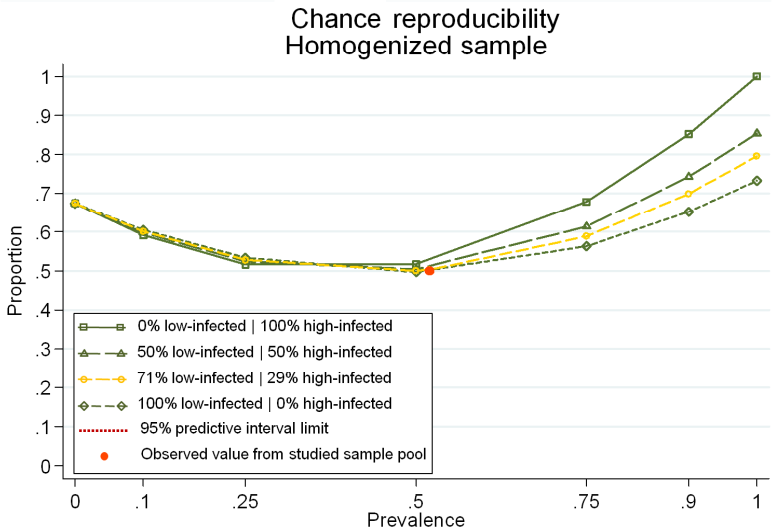
APPENDIX 11

Using Lenient-PGS definition for fish classification, these graphs predict descriptive chance repeatability of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and reproducibility for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.



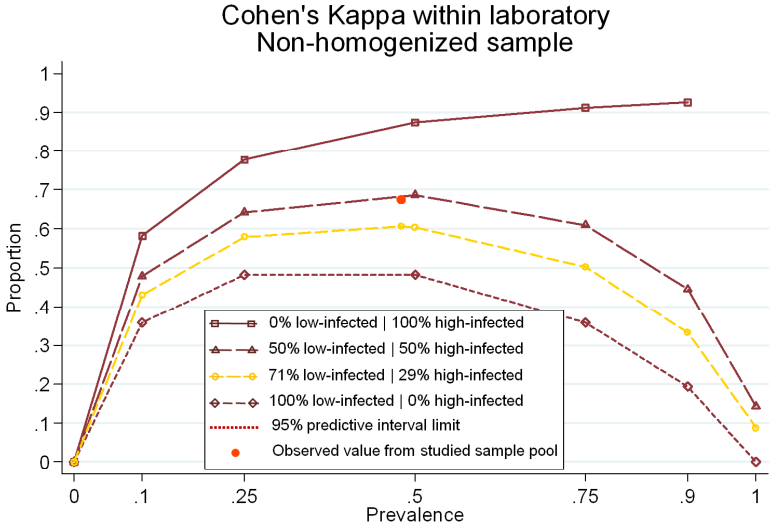




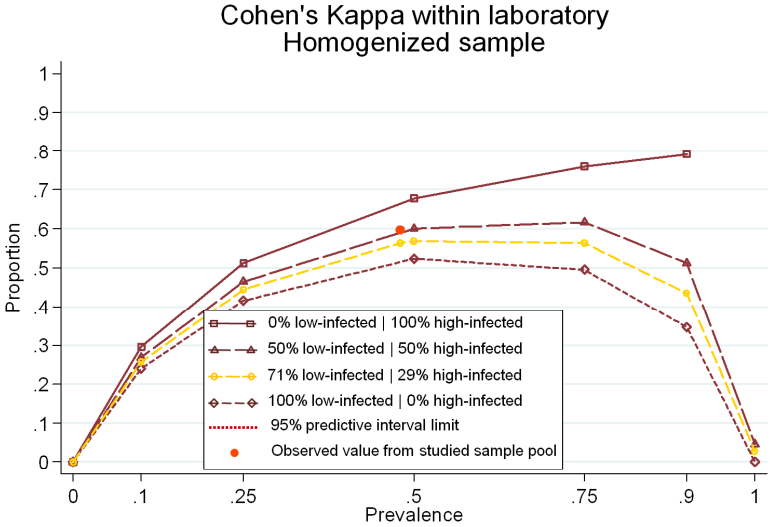


APPENDIX 12

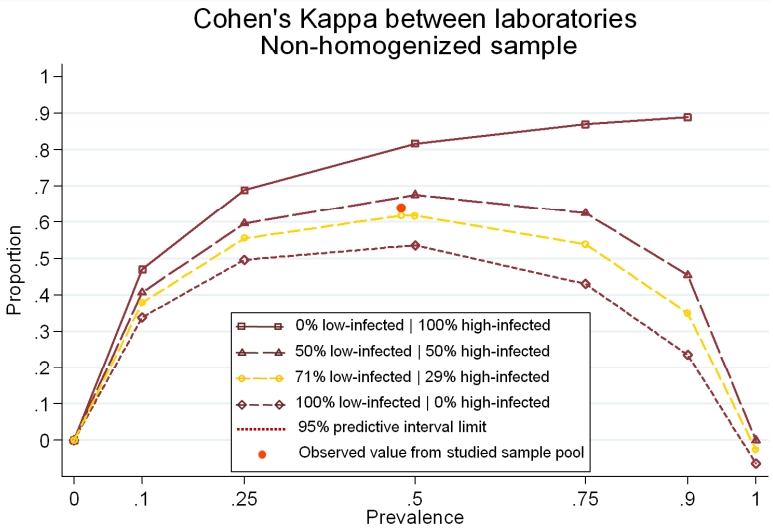
Using PGS definition for fish classification, these graphs predict descriptive Cohen's Kappa values of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and among laboratories for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.

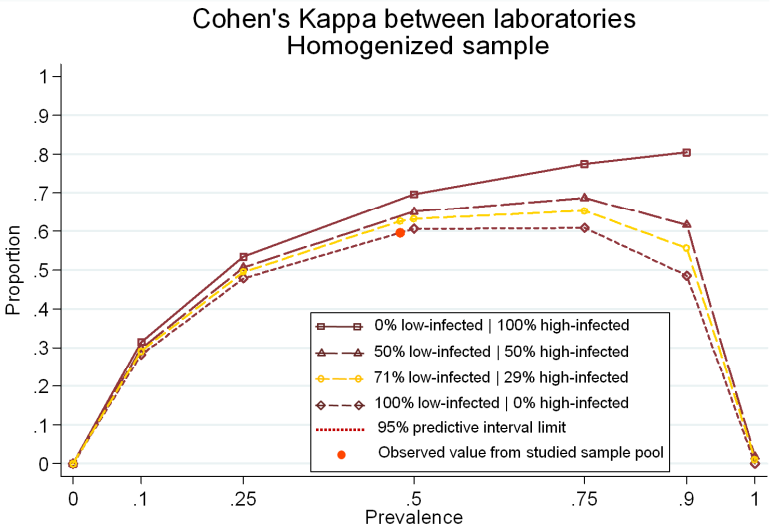


Graph. 5A



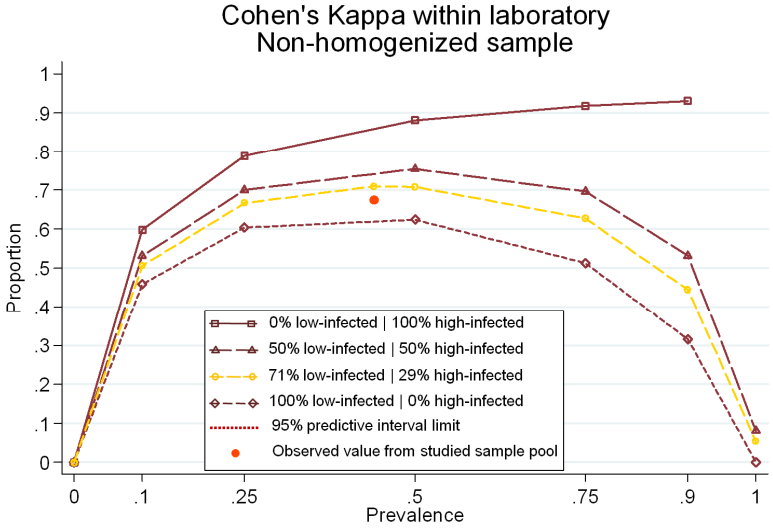
Graph. 5B

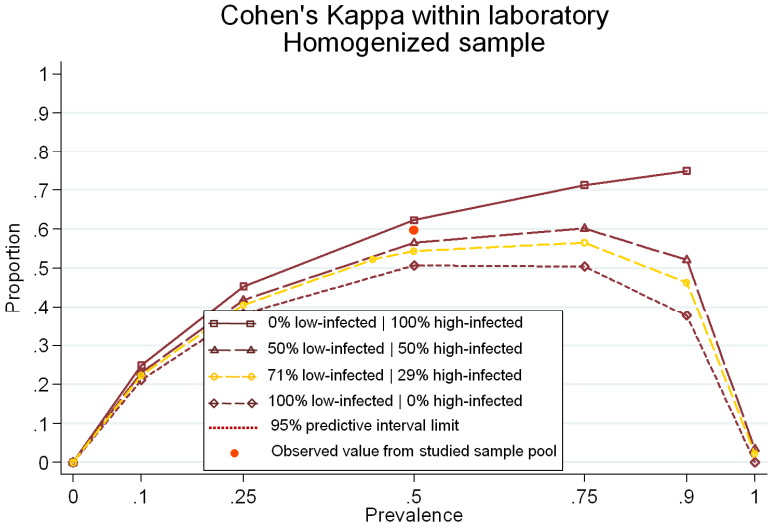


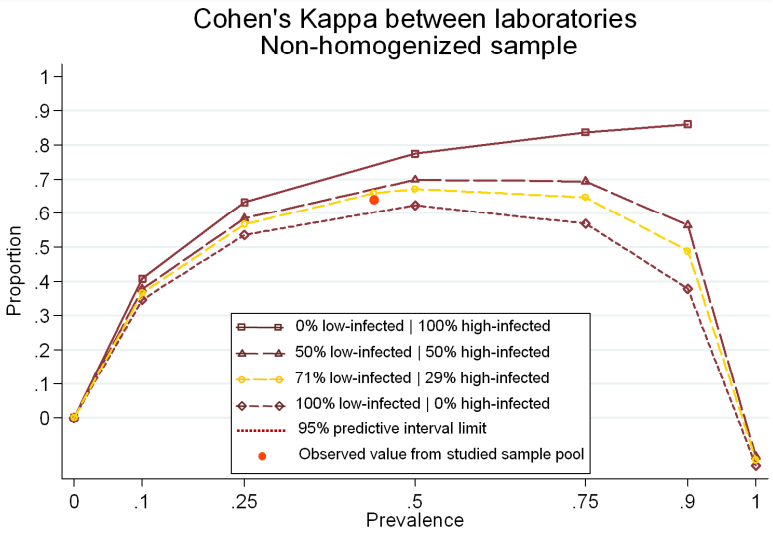


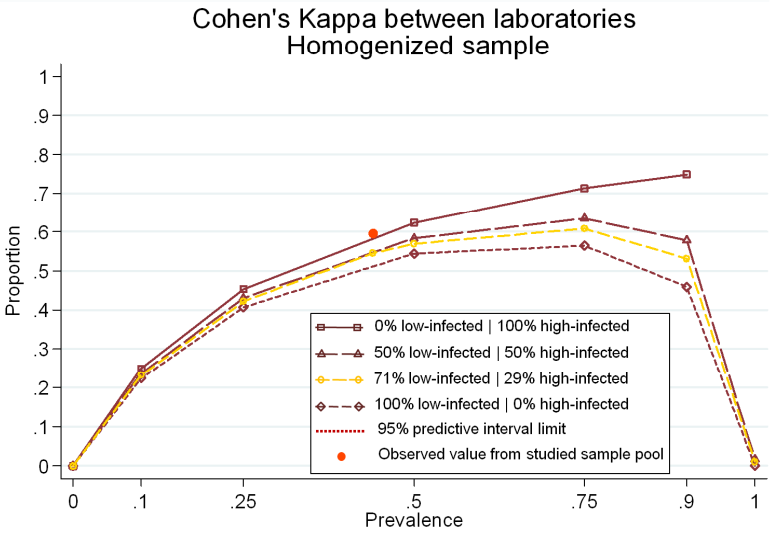
APPENDIX 13

Using Strict-PGS definition for fish classification, these graphs predict descriptive Cohen's Kappa values of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and among laboratories for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.



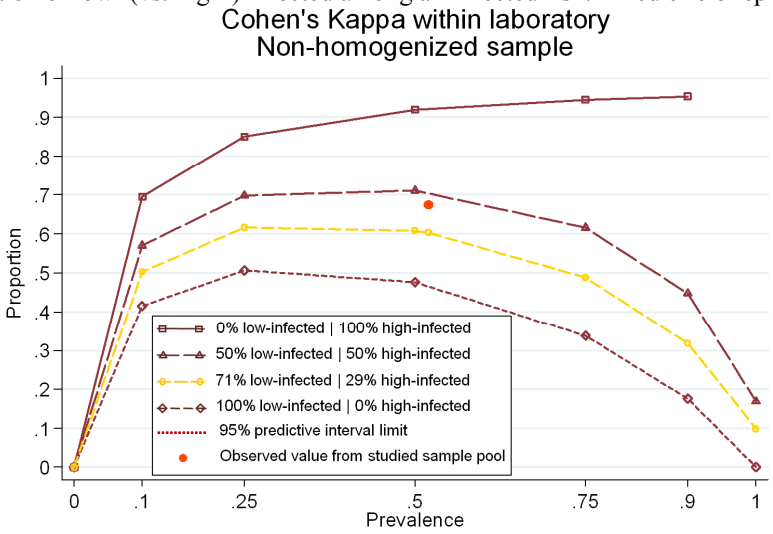


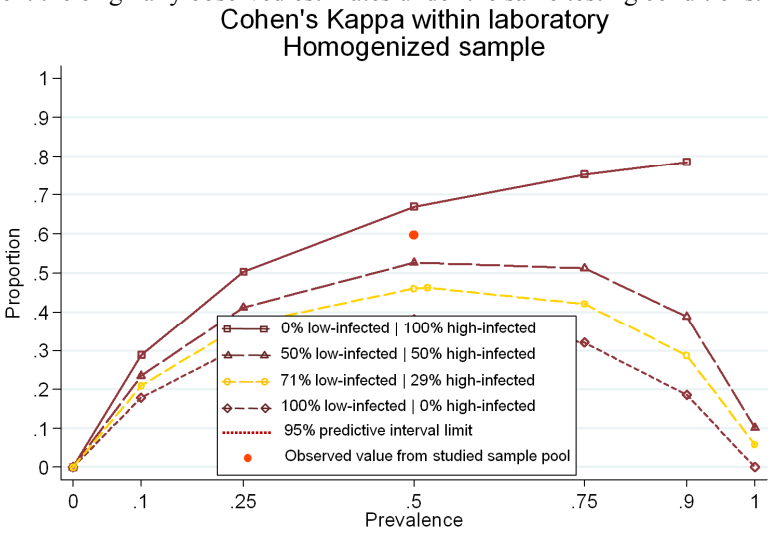


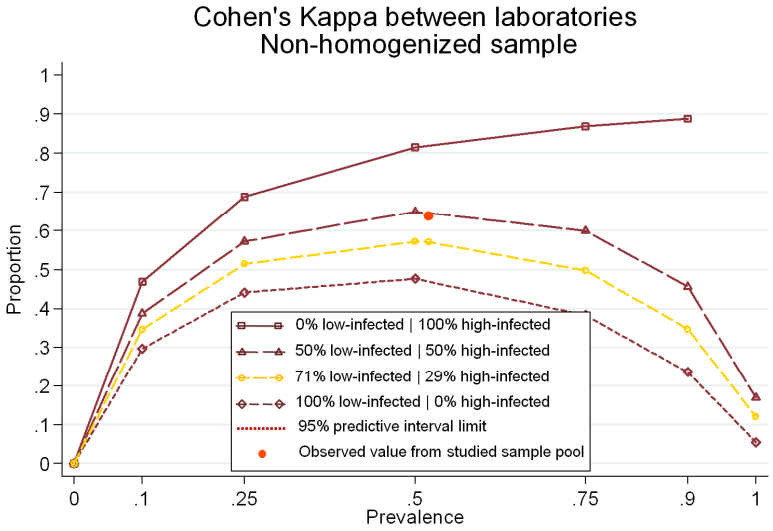


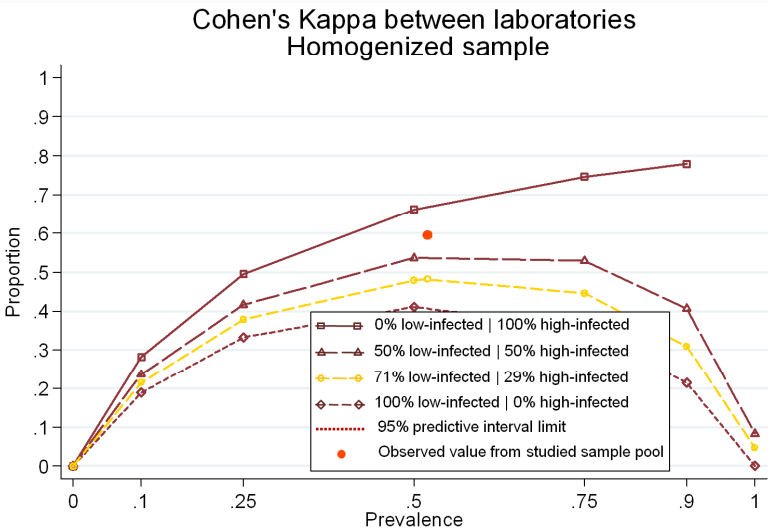
APPENDIX 14

Using Lenient-PGS definition for fish classification, these graphs predict descriptive Cohen's Kappa values of ISAV RT-PCR within the reference laboratory for non-homogenized and homogenized sample ; and among laboratories for non-homogenized and homogenized sample as a function of prevalence of infection and proportion of low- (vs. high-) infected among all infected fish. Filled circle represent the originally observed estimates under the same testing conditions.









APPENDIX 15

Validity of the assumption of constant DSe and DSp to predict agreement across prevalences

Prediction of agreement assumed that agreement within each infection categories was constant and that test performances did not depend on the factors varying across populations. Although, it is commonly accepted that DSp (or DSe) are constant parameters within each infection category and do not vary with population factors such as prevalence, various studies have challenged this postulate (e.g. Leeflang et al., 2009). Indeed, it is reasonable to think that when the prevalence of infection increases in a population, the proportion of advanced stages of infection also increases (Greiner & Gardner, 2000). It is also reasonable to assume that DSe will be greater in advanced stages compared to earlier stage of infection (Begg, 1987). Therefore, DSe and agreement in infected individuals may vary across prevalences. Furthermore, it is reasonable to think that test performance in NI individuals (DSp) is dependent on the pressure of cross-contamination from infected samples or/and positive controls. As the proportion of infected in the sampled population (prevalence) and/or the tested sample pool becomes larger, the pressure and chance of cross-contamination likely becomes greater. Therefore, DSp and agreement in non-infected individuals may also vary across prevalence groups.

The validity of these assumptions may introduce estimation bias for predicted agreement. For instance, agreement predicted for only NI fish (0% prevalence) may have been underestimated since agreement estimation within NI fish derived from a sample pool of 48% assumed prevalence corresponding to a much stronger pressure of cross-contamination. For instance, during a screening program, prevalence and associated contamination pressure are expected to be low and subsequently DSp and agreement to be higher. Assumption of constant performance within infection categories might not be appropriate and requires further investigation to estimate the degree of dependence between agreement and prevalence. If strong dependence of DSe and DSp with prevalence is confirmed, variation of DSe and DSp can be predicted from prevalence as illustrated by Brenner & Gefeller (1997). For instance, DSe and

DSp could be assessed for a 50% prevalence population and predicted across prevalences using information about population infection distribution (e.g. normal, bimodal). The constructed equation between DSe (or DSp) and prevalence can thereafter be directly included in Eqs. (2), (5) and (6). However, no information on the spectrum of ISAV loads across prevalences is yet available. By separating LI and HI fish, we intended to alleviate the dependence of DSe on prevalence, conditional on the assumption of constant DSe within each of the two sub-categories of infected salmon. Conversely, no sub-classification of NI fish was achievable and dependence of DSp on prevalence would have to be further investigated to refine the prediction of agreement in this infection category.

References:

Begg, C.B., 1987. Biases in the assessment of diagnostic tests. Stat. Med. 6, 411-423.

Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat. Med. 16, 981-991.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 42, 2-22.

Leeflang, M.M., Bossuyt, P.M., Irwig, L., 2009. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J. Clin. Epidemiol. 62, 5-12.

APPENDIX 16

Count per combination of test result of the five studied assays in each of the 4 prevalence level populations (low, mild, moderate, high). Studied assays included in the order: reverse-transcriptase polymerase chain reaction (RT-PCR), real-time RT-PCR (qRT-PCR), virus isolation (VI), indirect fluorescent antibody test (IFAT), lateral flow immunoassay (LFI).

Test result combination *(RT-PCR, qRT-PCR, VI, IFAT, LFI)*	Low Prevalence	Mild Prevalence	Moderate Prevalence	High Prevalence	Total
11111	0	0	0	31	31
11110	0	0	0	0	0
11101	0	0	0	6	6
11100	0	0	0	10	10
11011	0	0	0	5	5
11010	0	0	1	0	1
11001	0	0	0	0	0
11000	0	0	35	4	39
10111	0	0	0	0	0
10110	0	0	0	0	0
10101	0	0	0	0	0
10100	0	0	0	1	1
10011	0	0	0	0	0
10010	0	0	0	0	0
10001	0	0	0	0	0
10000	3	0	10	3	16
01111	0	0	0	0	0
01110	0	0	0	0	0
01101	0	0	0	0	0
01100	0	0	0	0	0
01011	0	0	0	0	0
01010	0	0	0	0	0
01001	0	0	0	0	0
01000	3	0	11	6	20
00111	0	0	0	0	0
00110	0	0	0	0	0
00101	0	0	0	0	0
00100	0	1	0	8	9
00011	0	0	0	0	0
00010	0	1	0	0	1
00001	0	0	0	0	0
00000	94	128	13	26	261
Total	**100**	**130**	**70**	**100**	**400**

APPENDIX 17

WinBUGS code for fitting a three-latent classes model with 4 populations and 5 tests with dependence among two pairs of tests (1&2 and 3&4), including the computation of the Bayesian p-value.

```
model
{
# prior distributions

selo[1] ~ dbeta(1,1);
selo[2] ~ dbeta(1,1);
selo[3] ~ dbeta(1,1);
selo[4] ~ dbeta(1,1);
selo[5] ~ dbeta(1,1);

sehi[1] ~ dbeta(1,1);
sehi[2] ~ dbeta(1,1);
sehi[3] ~ dbeta(1,1);
sehi[4] ~ dbeta(1,1);
sehi[5] ~ dbeta(1,1);

sp[1] ~ dbeta(1,1);
sp[2] ~ dbeta(1,1);
sp[3] ~ dbeta(1,1);
sp[4] ~ dbeta(1,1);
sp[5] ~ dbeta(1,1);


# population 1, informative prior set in data: alpha1=c(4.8,.6,.6)
for (i in 1:3) {
    pdir[1, i] <- delta[1, i] / sum(delta[1,])
    delta[1, i] ~ dgamma(alpha1[i], 1)
    }
  plo[1] <- pdir[1,2]
  phi[1] <- pdir[1,3]

# population 2, non-informative prior set in data: alpha2=c(1,1,1)
for (i in 1:3) {
    pdir[2, i] <- delta[2, i] / sum(delta[2,])
    delta[2, i] ~ dgamma(alpha2[i], 1)
    }
  plo[2] <- pdir[2,2]
  phi[2] <- pdir[2,3]

# population 3, non-informative prior set in data: alpha3=c(1,1,1)
for (i in 1:3) {
    pdir[3, i] <- delta[3, i] / sum(delta[3,])
    delta[3, i] ~ dgamma(alpha3[i], 1)
    }
  plo[3] <- pdir[3,2]
  phi[3] <- pdir[3,3]

# population 4, non-informative prior set in data: alpha4=c(1,1,1)
for (i in 1:3) {
    pdir[4, i] <- delta[4, i] / sum(delta[4,])
    delta[4, i] ~ dgamma(alpha4[i], 1)
    }
  plo[4] <- pdir[4,2]
  phi[4] <- pdir[4,3]



# define range for conditional covariances between test (1,2) and (4,5)

covlo12_low[1] <- max(-(1-selo[1])*(1-selo[2]),-selo[1]*selo[2]);
covlo12_upp[1] <- min(selo[1]*(1-selo[2]),(1-selo[1])*selo[2]);
covhi12_low[1] <- max(-(1-sehi[1])*(1-sehi[2]),-sehi[1]*sehi[2]);
covhi12_upp[1] <- min(sehi[1]*(1-sehi[2]),(1-sehi[1])*sehi[2]);
cov12_low[2] <- max(-(1-sp[1])*(1-sp[2]),-sp[1]*sp[2]);
cov12_upp[2] <- min(sp[1]*(1-sp[2]),(1-sp[1])*sp[2]);

covlo45_low[1] <- max(-(1-selo[4])*(1-selo[5]),-selo[4]*selo[5]);
covlo45_upp[1] <- min(selo[4]*(1-selo[5]),(1-selo[4])*selo[5]);
covhi45_low[1] <- max(-(1-sehi[4])*(1-sehi[5]),-sehi[4]*sehi[5]);
covhi45_upp[1] <- min(sehi[4]*(1-sehi[5]),(1-sehi[4])*sehi[5]);
cov45_low[2] <- max(-(1-sp[4])*(1-sp[5]),-sp[4]*sp[5]);
cov45_upp[2] <- min(sp[4]*(1-sp[5]),(1-sp[4])*sp[5]);


# define conditional covariance in infected: covlo12[1] ~ sensitivity low
infected, covhi12[1] ~ sensitivity high infected, covlo45[1] ~ sensitivity low
infected, covhi45[1] ~ sensitivity high infected

covlo12[1] ~ dunif(covlo12_low[1],covlo12_upp[1]);
covhi12[1] ~ dunif(covhi12_low[1],covhi12_upp[1]);
covlo45[1] ~ dunif(covlo45_low[1],covlo45_upp[1]);
covhi45[1] ~ dunif(covhi45_low[1],covhi45_upp[1]);


# define conditional covariance in non-infected:  cov12[2] ~ specificity,
cov45[2] ~ specificity

cov12[2] ~ dunif(cov12_low[2],cov12_upp[2]);
cov45[2] ~ dunif(cov45_low[2],cov45_upp[2]);


# multinomial models
for (k in 1:K){
  y[k,1:2,1:2,1:2,1:2,1:2] ~ dmulti(p[k,1:2,1:2,1:2,1:2,1:2], n[k])
  }
# cell probabilities expressed in terms of selo, sehi, sp, covlo, covhi, cov, plo
and phi
for (k in 1:K){

  p[k,1,1,1,1,1] <-
plo[k]*(selo[1]*selo[2]+covlo12[1])*selo[3]*(selo[4]*selo[5]+covlo45[1]) +
phi[k]*(sehi[1]*sehi[2]+covhi12[1])*sehi[3]*(sehi[4]*sehi[5]+covhi45[1]) + (1-
plo[k]-phi[k])*((1-sp[1])*(1-sp[2])+cov12[2])*(1-sp[3])*((1-sp[4])*(1-
sp[5])+cov45[2]);

  p[k,1,1,2,1,1] <- plo[k]*(selo[1]*selo[2]+covlo12[1])*(1-
selo[3])*(selo[4]*selo[5]+covlo45[1]) + phi[k]*(sehi[1]*sehi[2]+covhi12[1])*(1-
sehi[3])*(sehi[4]*sehi[5]+covhi45[1]) + (1-plo[k]-phi[k])*((1-sp[1])*(1-
sp[2])+cov12[2])*sp[3]*((1-sp[4])*(1-sp[5])+cov45[2]);

  p[k,1,2,1,1,1] <- plo[k]*(selo[1]*(1-selo[2])-
covlo12[1])*selo[3]*(selo[4]*selo[5]+covlo45[1]) + phi[k]*(sehi[1]*(1-sehi[2])-
covhi12[1])*sehi[3]*(sehi[4]*sehi[5]+covhi45[1]) + (1-plo[k]-phi[k])*((1-
sp[1])*sp[2]-cov12[2])*(1-sp[3])*((1-sp[4])*(1-sp[5])+cov45[2]);

  p[k,1,2,2,1,1] <- plo[k]*(selo[1]*(1-selo[2])-covlo12[1])*(1-
selo[3])*(selo[4]*selo[5]+covlo45[1]) + phi[k]*(sehi[1]*(1-sehi[2])-
covhi12[1])*(1-sehi[3])*(sehi[4]*sehi[5]+covhi45[1]) + (1-plo[k]-phi[k])*((1-
sp[1])*sp[2]-cov12[2])*sp[3]*((1-sp[4])*(1-sp[5])+cov45[2]);

  p[k,2,1,1,1,1] <- plo[k]*((1-selo[1])*selo[2]-
covlo12[1])*selo[3]*(selo[4]*selo[5]+covlo45[1]) + phi[k]*((1-sehi[1])*sehi[2]-
covhi12[1])*sehi[3]*(sehi[4]*sehi[5]+covhi45[1]) + (1-plo[k]-phi[k])*(sp[1]*(1-
sp[2])-cov12[2])*(1-sp[3])*((1-sp[4])*(1-sp[5])+cov45[2]);
```

```
  p[k,2,1,2,1,1] <- plo[k]*((1-selo[1])*selo[2]-covlo12[1])*(1-
selo[3])*(selo[4]*selo[5]+covlo45[1]) + phi[k]*((1-sehi[1])*sehi[2]-
covhi12[1])*(1-sehi[3])*(sehi[4]*sehi[5]+covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*(1-sp[2])-cov12[2])*sp[3]*((1-sp[4])*(1-sp[5])+cov45[2]);

  p[k,2,2,1,1,1] <- plo[k]*((1-selo[1])*(1-
selo[2])+covlo12[1])*selo[3]*(selo[4]*selo[5]+covlo45[1]) + phi[k]*((1-
sehi[1])*(1-sehi[2])+covhi12[1])*sehi[3]*(sehi[4]*sehi[5]+covhi45[1]) + (1-
plo[k]-phi[k])*(sp[1]*sp[2]+cov12[2])*(1-sp[3])*((1-sp[4])*(1-sp[5])+cov45[2]);

  p[k,2,2,2,1,1] <- plo[k]*((1-selo[1])*(1-selo[2])+covlo12[1])*(1-
selo[3])*(selo[4]*selo[5]+covlo45[1]) + phi[k]*((1-sehi[1])*(1-
sehi[2])+covhi12[1])*(1-sehi[3])*(sehi[4]*sehi[5]+covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*sp[2]+cov12[2])*sp[3]*((1-sp[4])*(1-sp[5])+cov45[2]);

  p[k,1,1,1,2,1] <- plo[k]*(selo[1]*selo[2]+covlo12[1])*selo[3]*((1-
selo[4])*selo[5]-covlo45[1]) + phi[k]*(sehi[1]*sehi[2]+covhi12[1])*sehi[3]*((1-
sehi[4])*sehi[5]-covhi45[1]) + (1-plo[k]-phi[k])*((1-sp[1])*(1-
sp[2])+cov12[2])*(1-sp[3])*(sp[4]*(1-sp[5])-cov45[2]);

  p[k,1,1,2,2,1] <- plo[k]*(selo[1]*selo[2]+covlo12[1])*(1-selo[3])*((1-
selo[4])*selo[5]-covlo45[1]) + phi[k]*(sehi[1]*sehi[2]+covhi12[1])*(1-
sehi[3])*((1-sehi[4])*sehi[5]-covhi45[1]) + (1-plo[k]-phi[k])*((1-sp[1])*(1-
sp[2])+cov12[2])*sp[3]*(sp[4]*(1-sp[5])-cov45[2]);

  p[k,1,2,1,2,1] <- plo[k]*(selo[1]*(1-selo[2])-covlo12[1])*selo[3]*((1-
selo[4])*selo[5]-covlo45[1]) + phi[k]*(sehi[1]*(1-sehi[2])-
covhi12[1])*sehi[3]*((1-sehi[4])*sehi[5]-covhi45[1]) + (1-plo[k]-phi[k])*((1-
sp[1])*sp[2]-cov12[2])*(1-sp[3])*(sp[4]*(1-sp[5])-cov45[2]);

  p[k,1,2,2,2,1] <- plo[k]*(selo[1]*(1-selo[2])-covlo12[1])*(1-selo[3])*((1-
selo[4])*selo[5]-covlo45[1]) + phi[k]*(sehi[1]*(1-sehi[2])-covhi12[1])*(1-
sehi[3])*((1-sehi[4])*sehi[5]-covhi45[1]) + (1-plo[k]-phi[k])*((1-sp[1])*sp[2]-
cov12[2])*sp[3]*(sp[4]*(1-sp[5])-cov45[2]);

  p[k,2,1,1,2,1] <- plo[k]*((1-selo[1])*selo[2]-covlo12[1])*selo[3]*((1-
selo[4])*selo[5]-covlo45[1]) + phi[k]*((1-sehi[1])*sehi[2]-
covhi12[1])*sehi[3]*((1-sehi[4])*sehi[5]-covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*(1-sp[2])-cov12[2])*(1-sp[3])*(sp[4]*(1-sp[5])-cov45[2]);

  p[k,2,1,2,2,1] <- plo[k]*((1-selo[1])*selo[2]-covlo12[1])*(1-selo[3])*((1-
selo[4])*selo[5]-covlo45[1]) + phi[k]*((1-sehi[1])*sehi[2]-covhi12[1])*(1-
sehi[3])*((1-sehi[4])*sehi[5]-covhi45[1]) + (1-plo[k]-phi[k])*(sp[1]*(1-
sp[2])-cov12[2])*sp[3]*((1-sp[4])*sp[5]-cov45[2]);

  p[k,2,2,1,2,1] <- plo[k]*((1-selo[1])*(1-selo[2])+covlo12[1])*selo[3]*((1-
selo[4])*selo[5]-covlo45[1]) + phi[k]*((1-sehi[1])*(1-
sehi[2])+covhi12[1])*sehi[3]*((1-sehi[4])*sehi[5]-covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*sp[2]+cov12[2])*sp[3]*((1-sp[4])*(1-sp[5])+cov45[2]);

  p[k,2,2,2,1,1] <- plo[k]*((1-selo[1])*(1-selo[2])+covlo12[1])*(1-
selo[3])*(selo[4]*selo[5]+covlo45[1]) + phi[k]*((1-sehi[1])*(1-
sehi[2])+covhi12[1])*(1-sehi[3])*(sehi[4]*sehi[5]+covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*sp[2]+cov12[2])*sp[3]*((1-sp[4])*(1-sp[5])+cov45[2]);

  p[k,1,1,1,1,2] <- plo[k]*(selo[1]*selo[2]+covlo12[1])*selo[3]*(selo[4]*(1-
selo[5])-covlo45[1]) + phi[k]*(sehi[1]*sehi[2]+covhi12[1])*sehi[3]*(sehi[4]*(1-
sehi[5])-covhi45[1]) + (1-plo[k]-phi[k])*((1-sp[1])*(1-sp[2])+cov12[2])*(1-
sp[3])*((1-sp[4])*sp[5]-cov45[2]);

  p[k,1,1,2,1,2] <- plo[k]*(selo[1]*selo[2]+covlo12[1])*(1-selo[3])*(selo[4]*(1-
selo[5])-covlo45[1]) + phi[k]*(sehi[1]*sehi[2]+covhi12[1])*(1-
sehi[3])*(sehi[4]*(1-sehi[5])-covhi45[1]) + (1-plo[k]-phi[k])*((1-sp[1])*(1-
sp[2])+cov12[2])*sp[3]*((1-sp[4])*sp[5]-cov45[2]);

  p[k,1,2,1,1,2] <- plo[k]*(selo[1]*(1-selo[2])-covlo12[1])*selo[3]*(selo[4]*(1-
selo[5])-covlo45[1]) + phi[k]*(sehi[1]*(1-sehi[2])-
covhi12[1])*sehi[3]*(sehi[4]*(1-sehi[5])-covhi45[1]) + (1-plo[k]-phi[k])*((1-
sp[1])*sp[2]-cov12[2])*(1-sp[3])*((1-sp[4])*sp[5]-cov45[2]);

  p[k,1,2,2,1,2] <- plo[k]*(selo[1]*(1-selo[2])-covlo12[1])*(1-
selo[3])*(selo[4]*(1-selo[5])-covlo45[1]) + phi[k]*(sehi[1]*(1-sehi[2])-
covhi12[1])*(1-sehi[3])*(sehi[4]*(1-sehi[5])-covhi45[1]) + (1-plo[k]-phi[k])*((1-
sp[1])*sp[2]-cov12[2])*sp[3]*((1-sp[4])*sp[5]-cov45[2]);

  p[k,2,1,1,1,2] <- plo[k]*((1-selo[1])*selo[2]-covlo12[1])*selo[3]*(selo[4]*(1-
selo[5])-covlo45[1]) + phi[k]*((1-sehi[1])*sehi[2]-
covhi12[1])*sehi[3]*(sehi[4]*(1-sehi[5])-covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*(1-sp[2])-cov12[2])*(1-sp[3])*((1-sp[4])*sp[5]-cov45[2]);

  p[k,2,1,2,1,2] <- plo[k]*((1-selo[1])*selo[2]-covlo12[1])*(1-
selo[3])*(selo[4]*(1-selo[5])-covlo45[1]) + phi[k]*((1-sehi[1])*sehi[2]-
covhi12[1])*(1-sehi[3])*(sehi[4]*(1-sehi[5])-covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*(1-sp[2])-cov12[2])*sp[3]*((1-sp[4])*sp[5]-cov45[2]);

  p[k,2,2,1,1,2] <- plo[k]*((1-selo[1])*(1-selo[2])+covlo12[1])*selo[3]*(selo[4]*(1-
selo[5])-covlo45[1]) + phi[k]*((1-sehi[1])*(1-sehi[2])+covhi12[1])*sehi[3]*(sehi[4]*(1-
sehi[5])-covhi45[1]) + (1-plo[k]-phi[k])*(sp[1]*sp[2]+cov12[2])*sp[3]*((1-sp[4])*sp[5]-
cov45[2]);

  p[k,2,2,2,1,2] <- plo[k]*((1-selo[1])*(1-selo[2])+covlo12[1])*(1-
selo[3])*(selo[4]*(1-selo[5])-covlo45[1]) + phi[k]*((1-sehi[1])*(1-
sehi[2])+covhi12[1])*(1-sehi[3])*(sehi[4]*(1-sehi[5])-covhi45[1]) + (1-
plo[k]-phi[k])*(sp[1]*sp[2]+cov12[2])*(1-sp[3])*((1-sp[4])*sp[5]-cov45[2]);

  p[k,1,2,2,2,2] <- plo[k]*(selo[1]*selo[2]+covlo12[1])*selo[3]*(selo[4]*(1-
selo[5])+covlo45[1]) + phi[k]*(sehi[1]*sehi[2]+covhi12[1])*sehi[3]*((1-
sehi[4])*(1-sehi[5])+covhi45[1]) + (1-plo[k]-phi[k])*((1-sp[1])*(1-
sp[2])+cov12[2])*(1-sp[3])*(sp[4]*sp[5]+cov45[2]);

  p[k,1,1,2,2,2] <- plo[k]*(selo[1]*selo[2]+covlo12[1])*(1-selo[3])*((1-
selo[4])*(1-selo[5])+covlo45[1]) + phi[k]*(sehi[1]*sehi[2]+covhi12[1])*(1-
sehi[3])*((1-sehi[4])*(1-sehi[5])+covhi45[1]) + (1-plo[k]-phi[k])*((1-sp[1])*(1-
sp[2])+cov12[2])*sp[3]*(sp[4]*sp[5]+cov45[2]);

  p[k,1,2,1,2,2] <- plo[k]*(selo[1]*(1-selo[2])-covlo12[1])*selo[3]*((1-
selo[4])*(1-selo[5])+covlo45[1]) + phi[k]*(sehi[1]*(1-sehi[2])-
covhi12[1])*sehi[3]*((1-sehi[4])*(1-sehi[5])+covhi45[1]) + (1-plo[k]-phi[k])*((1-
sp[1])*sp[2]-cov12[2])*(1-sp[3])*(sp[4]*sp[5]+cov45[2]);

  p[k,1,2,2,2,2] <- plo[k]*(selo[1]*(1-selo[2])-covlo12[1])*(1-selo[3])*((1-
selo[4])*(1-selo[5])+covlo45[1]) + phi[k]*(sehi[1]*(1-sehi[2])-covhi12[1])*(1-
sehi[3])*((1-sehi[4])*(1-sehi[5])+covhi45[1]) + (1-plo[k]-phi[k])*(sp[1]*(1-
sp[2])-cov12[2])*sp[3]*(sp[4]*sp[5]+cov45[2]);

  p[k,2,1,2,2,2] <- plo[k]*((1-selo[1])*selo[2]-covlo12[1])*selo[3]*((1-
selo[4])*(1-selo[5])+covlo45[1]) + phi[k]*((1-sehi[1])*sehi[2]-
covhi12[1])*sehi[3]*((1-sehi[4])*(1-sehi[5])+covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*(1-sp[2])-cov12[2])*(1-sp[3])*(sp[4]*sp[5]+cov45[2]);

  p[k,2,2,2,2,2] <- plo[k]*((1-selo[1])*(1-selo[2])+covlo12[1])*(1-selo[3])*((1-
selo[4])*(1-selo[5])+covlo45[1]) + phi[k]*((1-sehi[1])*(1-
```

```
sehi[3])*((1-sehi[4])*(1-sehi[5])+covhi45[1]) + (1-plo[k]-
phi[k])*(sp[1]*sp[2]+cov12[2])*sp[3]*(sp[4]*sp[5]+cov45[2]);
}

# code for Bayesian p-value
for (k in 1:K){
  s[k,1:2,1:2,1:2,1:2,1:2] ~ dmulti(p[k,1:2,1:2,1:2,1:2,1:2],n[k])
  for (i1 in 1:2){
    for (i2 in 1:2){
      for (i3 in 1:2){
        for (i4 in 1:2){
          for (i5 in 1:2){
        x2[k,i1,i2,i3,i4,i5] <- pow(y[k,i1,i2,i3,i4,i5]-
p[k,i1,i2,i3,i4,i5]*n[k],2)/(p[k,i1,i2,i3,i4,i5]*n[k])
        ref[k,i1,i2,i3,i4,i5] <- pow(s[k,i1,i2,i3,i4,i5]-
p[k,i1,i2,i3,i4,i5]*n[k],2)/(p[k,i1,i2,i3,i4,i5]*n[k])
  }}}}}}
x2total <- sum(x2[,,,,,])
reftotal <- sum(ref[,,,,,])
bayesp <- step(reftotal-x2total)
}

# note: required order of cell counts:
#     (RT-PCR=+,qRT-PCR=+,VI=+,IFAT=+,KIT=+), (+,+,+,+,-), (+,+,+,-,+),(+,+,+,-,-),
(+,+,-,+,+), (+,+,-,+,-), (+,+,-,-,+), (+,+,-,-,-), (+,-,+,+,+), (+,-,+,+,-), (+,-
,+,-,+), (+,-,+,-,-), (+,-,-,+,+), (+,-,-,+,-), (+,-,-,-,+), (+,-,-,-,-),(-
,+,+,+,+), (-,+,+,+,-), (-,+,+,-,+),(-,+,+,-,-), (-,+,-,+,+), (-,+,-,+,-), (-,+,-,-
,+), (-,+,-,-,-), (-,-,+,+,+), (-,-,+,+,-), (-,-,+,-,+), (-,-,+,-,-), (-,-,-,+,+),
(-,-,-,+,-), (-,-,-,-,+), (-,-,-,-,-)
#     for popul 1,2,...
# note: first number in c() list for .Dim is the number of populations (K)
list(K=4, n=c(100,130,70,100), alpha1=c(4.8,.6,.6), alpha2=c(1,1,1),
alpha3=c(1,1,1), alpha4=c(1,1,1),
y=structure(.Data=c(0,0,0,0,0,0,0,0,0,0,0,0,0,3,0,0,0,0,0,0,3,0,0,0,0,0,0,9
4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,128,0,0,0,0,0,1,0,3
5,0,0,0,0,0,0,0,0,10,0,0,0,0,0,0,0,11,0,0,0,1,0,1,0,13,31,0,6,10,5,0,0,4,0,0,0,1,0,0,
0,3,0,0,0,0,0,0,0,6,0,0,0,8,0,0,0,26),.Dim=c(4,2,2,2,2,2)))

# note: one initial value for each popul, in addition to the se/sp's
list(selo=c(0.5, 0.5, 0.5,0.5,0.5), sehi=c(0.7, 0.7, 0.7,0.7,0.7), sp=c(0.9, 0.9,
0.9,0.9,0.9), covhi12(0), covhi45(0), covlo12(0))

list(selo=c(0.6, 0.6, 0.1,0.1,0.1), sehi=c(0.8, 0.8, 0.8,0.7,0.7), sp=c(0.95, 0.95,
0.95,0.95,0.95), covhi12(0.05), covhi45(0.15), covlo12(0.1),
delta=structure(.Data=c(1,0.02,0.01,0.6,0.4,0.01,0.3,0.1,0.6),.Dim=c(3,3))

list(selo=c(0.4, 0.4, 0.1,0.1,0.1), sehi=c(0.8, 0.8, 0.8,0.8,0.8), sp=c(0.9, 0.9,
0.9,0.9,0.9), covhi12(0), covhi45(0), covlo12(0),
delta=structure(.Data=c(1,0.02,0.01,0.4,0.6,0.01,0.1,0.1,0.8),.Dim=c(3,3))
```